ALMA MATER STUDIORUM A.D. 1088
UNIVERSITÀ DI BOLOGNA

A network analysis of the California road map

Guido Campani

Department of Physics and Astronomy
ALMA MATER STUDIORUM- UNIVERSITÀ DI BOLOGNA
Bologna, Italy 2021

A network analysis of the California road map
Report complex network course
Guido Campani

Professor: Daniele Remondini

Report complex network course 2021
Department of Physics and Astronomy
Alma Mater Studiorum- Università di Bologna

Cover: A representation of a reduced number of nodes of the California Road map. The size of the node and the links grow with the nodes and links betweeness centrality, while the color is related to the degree of them.

Contents

List of Figures

List of Tables

1
Introduction

In this project a map of the Californian road will be analyzed. The graph taken into account was downloaded from http://snap.stanford.edu/data/roadNet-CA.html. Due to the dimension of the whole network, the graph is composed by 1965206 nodes and 2766608, just a smaller portion of it will be considered (18000 nodes and 26763 edges). The features taken into account in order to measure the network will be the nodes degree, their betweeness and closeness centrality and the nodes clustering coefficient. As a first step it will be verified any size invariant of the system. Secondly it will be determined, where it is possible, from which there is a qualitative and or quantitative equivalence in the distributions of the four features mentioned before. No evaluation of the equivalence of different sub-graph of the same size will be made. Even if it is an interesting step for a more general validation of the report results, it won't be made because only this part would be a more burdensome work then the achieving of all the other results its self. So the results of the following steps have to be intended related only to the reduced graph taken into account. So, after the general analysis of the whole graph the reduced network is analyzed more deeply. His features are compared to the ones of an Erdős–Rényi (E-R) graph in order to underline which are characteristic due the his structure or more related to a random factor. Furthermore both the road map and the E-R are perturbed in different way. This method allow to investigate the presence of any relevance of nodes with peculiar characteristic. In the end of the project a building network algorithm is presented and analyzed. This algorithm, starts from the position of each node and try to copy the real map relying on the degree distribution of the nodes and on the linkage probability in function of their distance.

2

Theory

This chapter presents the section levels that can be used in the template.

2.1 Network basic theory

Most of the theory used to analyze the data are based on few concept.
Let's define a network as a set of elements(N) that have some linking relations among them. Each element is called node. If two node are linked together there will be an edge. Essentially, a graph G= (V, L) is a combination of these two sets; a set V of N nodes and a set L of E links connecting the corresponding nodes.
The next step is to define the adjacency matrix A (NxN), in which a_{ij}=1.If there is a link between vertices i and j, otherwise, a_{ij}=0. Please note in this report only unweighted and undirected graph are taken into account.
In order to classify the nodes in a graph is it important to study the connection among them and how to get from one to another. So a walk is the sequence of all nodes and edges between 2 nodes and a path is a walk with all distinct nodes and links. Instead a closed path is called cycle.
In this report a first analysis of the graph has been made using the definition of degree, centrality of a node, closeness centrality and clustering coefficient.
Degree
The degree k of th node v is the number of edges adjacent to it.

$$k(v) = \sum_{j \in V} a_{vj} \tag{2.1}$$

Betweeness centrality
Betweeness centrality of a node v is defined as:

$$B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{2.2}$$

Where V is the set of nodes, σ(s,t) is number of shortest path from s to t, and σ(s,t|v) is the number of those same paths which pass trough node v. If s=t, σ(s,t)=1 and if $\in s, t, \text{œ}(s,t|v) = 0$.

Closeness centrality
The closeness centrality of a node u is the reciprocal of the average shortest path distance to u over all n-1 reachable nodes weighted by the size of the component it

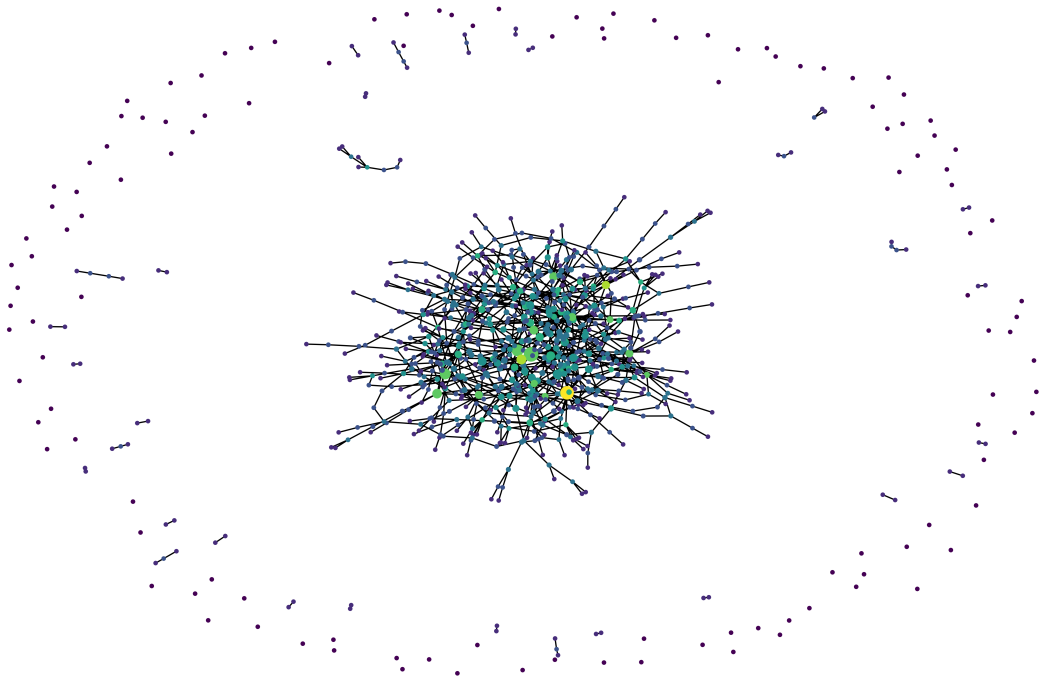


Figure 2.1: Erdős–Rényi network

belongs to.

$$C(u) = \frac{n-1}{N-1} \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)} \tag{2.3}$$

Clustering coefficient
For unweighted graphs, the clustering of a node u is the fraction of possible triangles through that node that exist,

$$c_u = \frac{2T(u)}{deg(u)deg(u)-1} \tag{2.4}$$

where T(u) is the number of triangles through node u and deg(u) is the degree of u.

2.1.1 Giant component

Another important feature to study in a network is the so called giant component.A "component" is a subgroup of the graph nodes in which any two vertices are connected to each other by paths, and which is connected to no additional vertices in the rest of the graph. A network has a "giant component", if it contains a finite fraction of the entire graph's vertices almost every node is reachable from almost every other.

2.1.2 Erdős–Rényi graph

The Erdős–Rényi graph is a kind of random graph in which the set of N node are linked together randomly. Starting from a set with no edges, all the possible couple

of nodes are taken into account and between each of them a linkage can be formed with a probability p. So given p, the number N of node and the total number of edges E, the probability to get a specified graph is it equal to:

$$P = p^E(1-p)^{\binom{N}{2}-E}$$

. For this reason it is also called random Bernoulli graph

The distribution of the degree of any particular vertex is binomial:

$$P(deg(v) = k) = \binom{N-1}{k} p^k (1-p)^{N-1-k} \tag{2.5}$$

as n $\to \infty$

$$P(deg(v) = k) = \frac{(Np)^k e^{-Np}}{k!} \tag{2.6}$$

Let's define λ=Np as the connectivity. In the limit of V»1 a ER graph tends to a regular graph with equal λ connectivity for all nodes.

- if λ<1 , then a graph in G(N, p) will almost surely have no connected components of size larger than O(log(N)).
- if λ=1, then a graph in G(N, p) will almost surely have a largest component whose size is of order $N^{2/3}$.
- if λ>1, then a graph in G(N, p) will almost surely have a unique giant component containing a positive fraction of the vertices. No other component will contain more than O(log(N)) vertices

In such a network for N increasing a phase transition of the giant component occurs. A random Bernoulli graph is very often used as null model in order to compare this one to an understudy graph. In these cases the main aim is to underline the presence of any structural property or if on the contrary some features is more randomic.

3
Methods

3.1 Data information and procedure

The graph taken into account was downloaded from http://snap.stanford.edu/data/roadNet-CA.html, It represents a road network of California. Intersections and endpoints are represented by nodes and the roads connecting these intersections or road endpoints are represented by undirected edges. The file is composed by couple of numbers which represents the edges. a The graph is formed by 1965206 nodes and 2766608 links. The procedure adopted for the analysis of the network was the following:

- Verifying of the possibility to exploit only a reduced part of the graph with compatible characteristic.
- Comparing of the graph with E-R graph
- Modeling of an algorithm to reproduce the network

3.2 Reduced network

Working with all the data is really time consuming, so in order to come up to some results in a reasonable time period was fundamental to use just a reduced network. This leads to at least three questions: is this equivalent to analyze the whole system? How to choose this subset of nodes? Is there a suitable number of nodes for the reduced graph for which some of the interested features have stable behaviour with the growth in size?

Answering the the first question may be very long and complex. But if one is interested just in only few features he can compare the ones of the whole graph with the ones of reduced one and see if they are compatible.

For this report in order to classify the network only the following distribution of the network was taken in account: degree, betweeness centrality, closeness centrality and clustering. So the second and third question concerned only these quantity.
At a first glance if only a small random set S of nodes of the whole network is taken into account ($|S| < 10^6$) there is no giant component. All the nodes are connected in infinitesimal group (compared to the total number of vertices) and the concerning features will vary a lot during the growth in number. So it may be not useful to proceed in this way because in order to have some stable behaviour it is necessary to reach a to high number of vertices and the analysis will be time demanding.

Since the data represents a road map, nodes will be linked following some topological rule, it would be very improbable that two nodes that are far away will be connected. Since the label of the node is related on their position in the space (two near points have no far label-number) the procedure used was simply to start from the first node of the data and to take L number of links. The links list had been sorted in ascending order taking in consideration the first element of the couple. This procedure allows to have at least one big component from the first nodes.

3.2.1 Size invariant of the main features

The first part of the analysis concerned the third question. Several networks were built following the second procedure and increasing for each step the number of links and so of nodes. For each steps the four main features distribution were measured and the compatibility of these distributions along the size increazing was studied. The compatibility taken into account was of two types: qualitatively and quantitatively. For both of them the data were divided in bins. For the qualitatively analysis the ranking of the bins for each distribution was compared and also the shape of the distributions itself (flat or pecked, the number and position of the pecks). For the quantitatively analysis the χ^2 test was performed each time in order to have a distance measures between two distributions. It is not expected a p-value of the test > 0.05 and so a compatibility among them, but just an information to evaluate how far distributions can get.
For the furthers analysis a reduced graph was chosen. The criteria used to cut the number of nodes of the network were the time used to calculate the four distributions and the proximity of these distributions to the ones of the whole graph based on the χ^2 test.

3.2.2 Comparing with the E-R graph

In order to verify if some of the characteristic of the subset were structural or due to some random phenomena an E-R, with the same size and linking probability of the reduced road map, was taken into account . For the first step a phase transition procedure was performed and the reduced road graphs and the E-R network behaviour were compared. For the procedure the links of the two networks were added in a random order, step by step starting from a zero links condition. Then the four distributions were compared. A qualitative analysis was also made dividing the networks into communities by using the *greedy_modularity_communities* function of the Networkx library. "This function uses Clauset-Newman-Moore greedy modularity maximization. Greedy modularity maximization begins with each node in its own community and joins the pair of communities that most increases modularity until no such pair exists".[1] Finally the two networks were perturbed removing nodes. At first the nodes were chosen following two different rules: the ones with the highest

[1] https://networkx.org/documentation

degree or the one with the highest betweeness centrality ,then they were picked up randomly. So the structural consequences on the two were studied.

3.2.3 Copycat network

The last aim of the project was trying to emulate the the reduced graph, in particular to reproduce the four features distributions. The approach used was to exploit the topological characteristics of data. From the initial graph the position of the nodes were obtained applying the function *spring_layout* of the library Networkx. "The function uses the Fruchterman-Reingold force-directed algorithm.It simulates a force-directed representation of the network treating edges as springs holding nodes close, while treating nodes as repelling objects, sometimes called an anti-gravity force. Simulation continues until the positions are close to an equilibrium." [2] Next the distance between all the nodes were calculated. In order to build a distribution of the distances the distances its self were binned in different ranges. The probability to have a link was calculated for each distance bin by counting the ratio between the number of linked nodes within the the distance range and the total number of nodes in the same distance range. Then starting from the same nodes' position a new graph was built. For each couple of nodes an edge was added with a probability distribution based on this distance procedure. Finally the following correction were added:

- Nodes with the highest degree were removed until the maximum degree of the new graph was equal to the one of the reduced network.
- Edges of nodes with degree equal to the maximum or the maximum-1 were removed until the the ratio of the nodes of the copycat network with degree equal to three is equal the the one of the reduced network, .
- A loop procedure was used to link isolated node with the connected ones. The algorithm chooses, if it was possible, the nodes that have a higher distance probability but also that re-balance the degree distribution in order to reach the same of the initial network.
- The algorithm removes with a while loop edges of nodes with degree equal to two if they are too much in order to re-balance the degree distribution.
- The algorithm adds with a while loop edges to nodes with degree equal to one if they are too much in order not to go out of nodes with degree two and to decrease node of degree one. The nodes chosen are the more distance-probable and with degree equal to two.
- The algorithm melts to the graph components smaller than three in order not to have to many small components.

[2] https://networkx.org/documentation

4

Results

4.1 Main features size compatibility

4.1.1 Betweeness centrality

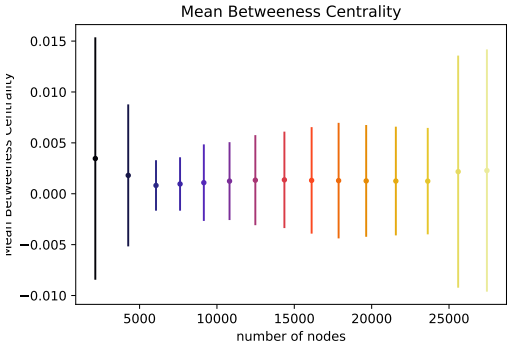
As figure 4.1a shown very clear, the mean value of the betweeness centrality is constant with the increases of the size. At least from 2139 nodes to 27449 nodes. All the distributions are very pecked around 0, as instance it is possible to give a look to the graph 4.10c to see qualitatively the shape of the distribution. For higher order of magnitude it was almost impossible run the code due to the high time-consume of the betweeness algorithm. Their cumulative distribution looks like very similar qualitatively speaking, 4.1b, but the χ^2 test of the distributions is for all of them higher then 23.6(where 23.6 is the maximum value to accept the compatibility hypotheses with a d.o.f. of 9, the distribution was divided in ten bins). The mean value of the χ^2 is 144 ± 121

4.1.2 Closeness centrality

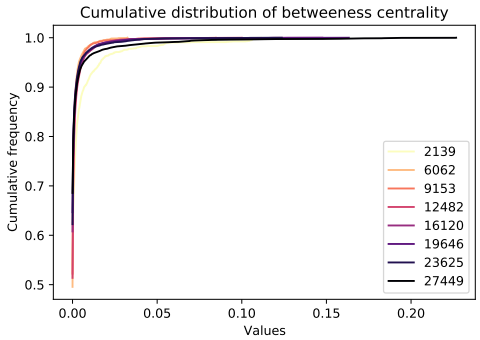
The measure of closeness centrality is strongly related to growth of the giant component. Indeed, as shown in 4.10e the distribution has two group of pecks: the left ones related to all the small components of the graph, and the right ones related to the giant component. As the cumulative distribution graph shows 4.2b with the increase of the number of nodes/links the left part of the distribution tends to disappear and the right peaks increases start melting together into a unique peak which gets thinner. So even if the mean values of the distributions with different number of nodes are compatible 4.2a the distributions are qualitatively different. It is possible to observe the left part of the distribution is still presents also with 41561 nodes 4.3.It was almost impossible run the code for higher size due to the high time-consume of the betweeness algorithm.

4.1.3 Clustering coefficient

As seen in figure 4.5 this distributions are qualitative similar. The average value is equal 0.046 ± 0.002. However, the *chi*2 test shows shows no compatibility among all the distributions with a mean χ^2 value of 247 ± 307. A χ^2 test compatibility occurs for size bigger than 1'567'683 with a value equal to $19 \pm 13 < 23.6$ (where 23.6 is the maximum value to accept the compatibility hypotheses with a d.o.f. of 9, the distribution was divided in ten bins).

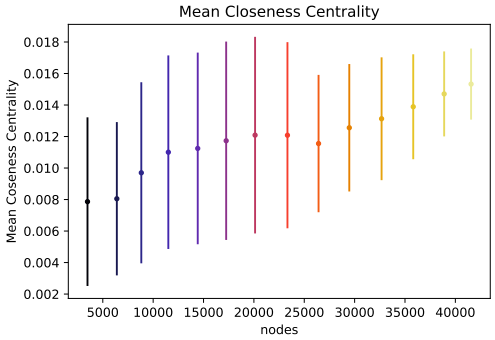


(a) *Mean betweeness for increasing number of nodes*

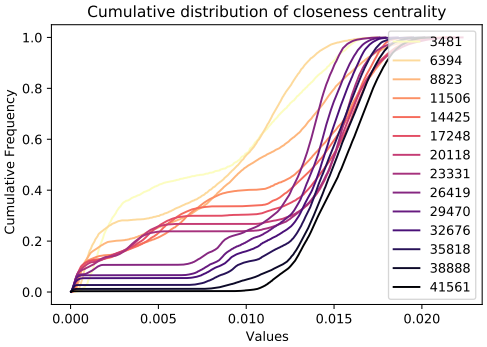


(b) *Cumulative distribution of betweeness for increasing number of nodes*

Figure 4.1: Betweeness for increasing number of nodes



(a) *Mean Closeness for increasing number of nodes*



(b) *Cumulative distribution of closeness for increasing number of nodes*

Figure 4.2: Closeness centrality for increasing number of nodes

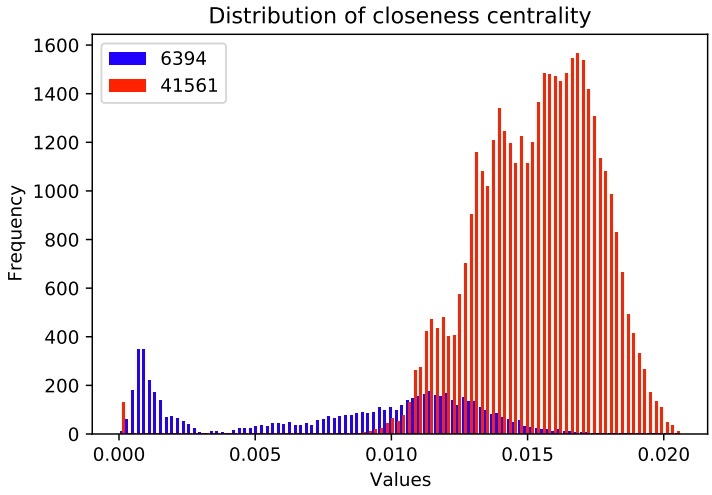


Figure 4.3: Distribution of closeness centrality for networks with size equal to 6'394 and 41'561

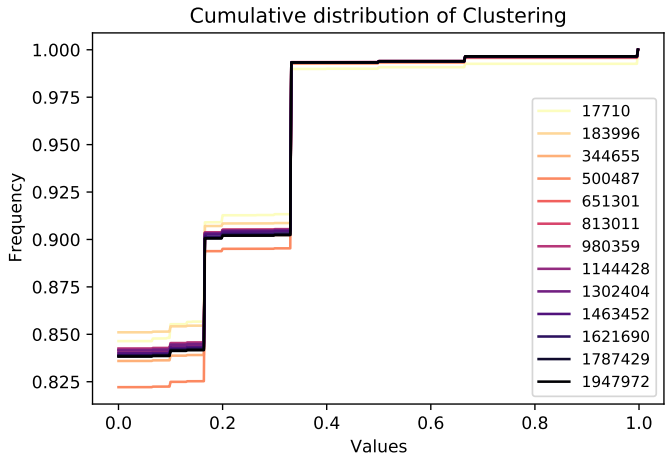


Figure 4.4: Clustering cumulative distribution for increasing number of nodes

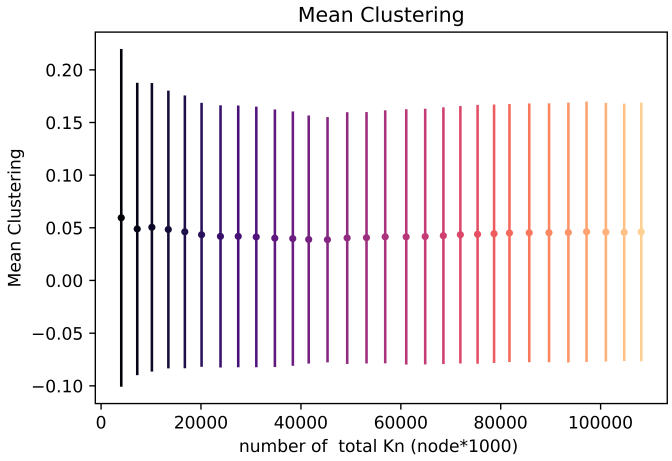


Figure 4.5: Mean clustering for increasing number of nodes

4.1.4 Degree

In figure 4.6 it is possible to see the distribution of the degree of the nodes of the whole system. The figure 4.7b shows the variation of the ratio increasing the number of edges with step of one thousand links. The figure 4.7a , which is a zoom of the first $4 \cdot 10^4$ nodes, shows that after around 7200 nodes the ranking of the nodes' degree is qualitatively the same (the same order). From step 9 (7236 nodes) till step 13 (9554 nodes) the ratio of the degrees 4 has an increase of 18 percentage points. in the same range the ratio of the degree 1 goes from 0.221 to 0.170, which is equal to a variation of 23 percentage points. From networks whose size is bigger then $1.8 \cdot 10^4$ the ratio variations start to be more smooth. If only the first five degree are taken into account, the average of these variation from the final values is in percentage 0.8%. Considering the χ^2 test the compatibility is reached after size=1'788'847.Indeed, from this size on the test shows a value of $15 \pm 13 < 28.3$ (where 28.3 is the maximum value to reject the hypothesis of compatibility). In conclusion this feature is qualitatively invariant after size equal to 7236, it conserves its order. After size $1.8 \cdot 10^4$ the variation are smooth, but can be considered quantitatively invariant after size 1'788'847 .

4.1.5 Choice of the reduced network size

Since the time demanding of the betweeness and closeness algorithm is very high, the size of the network is cut at 18'000 nodes. It is a number low enough to allows the code to run enough fast, but it is not so low to be qualitatively different from the whole system in relation with the degree, clustering and betweeness distributions. Unfortunately in order to have closeness centrality distribution qualitatively suitable it is necessary a too large set to made the code run fast. In order to have a quantitatively compatibility with the whole network for the degree and clustering distributions a proper size should be greater then $1.8 \cdot 10^6$ and it is not time feasible. For the closeness and betweeness it was impossible to find this threshold

4.2 Reduced network 18'000 nodes

4.2.1 Component and phase transition

The network doesn't have a proper giant component, as shown in table 4.1, the first 10 component represent the 90% of the population while the first one is the 64% of the population. While in the E-R graph the 94% of the popoluation belong to the giant component.

The random network, as expected has a phase transition of the ratio $\frac{|Giant|}{|G|}$ for $\lambda = 1$ while for the Road graph we has a transition around $\lambda = 1.6$

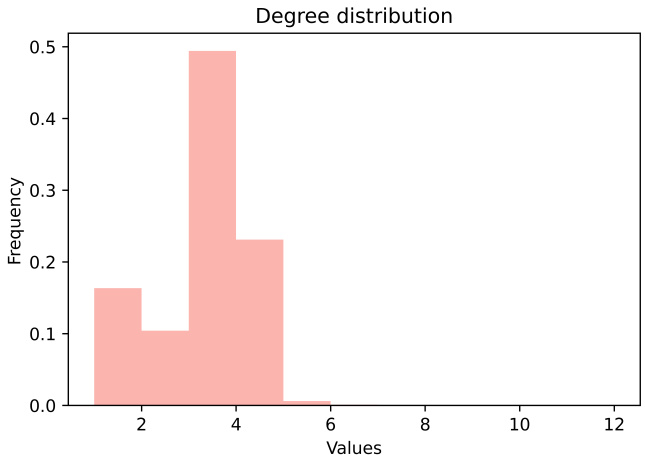
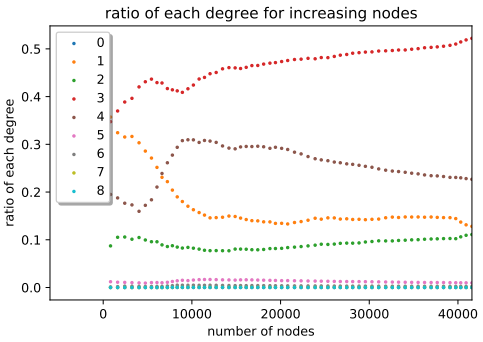
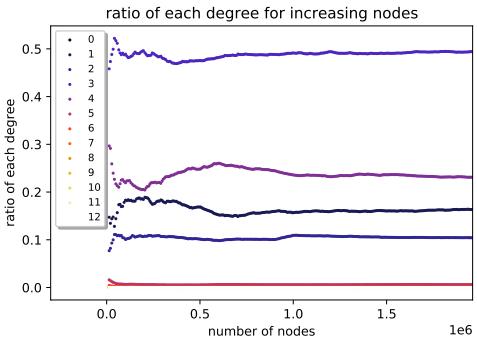


Figure 4.6: Degree distribution California road map with 1965206 nodes and 2766608 links



(a) *Frequency of nodes with the same degree for increasing number of networks nodes*

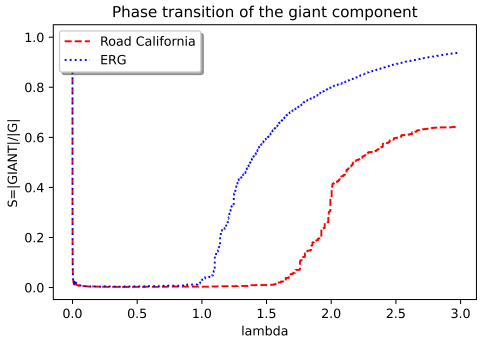


(b) *Frequency of nodes with the same degree for increasing number of networks nodes*

Figure 4.7: Degree convergence

n component	size
0	11556
1	2733
2	953
3	306
4	240
5	146
6	124
7	114
8	91
9	74

Table 4.1: Nummber of nodes for each component of a network with 18000 nodes sorted by length



(a)

Figure 4.8: Phase transition

$Degree$	$ratioRoad-map$	$ratioE-R$
0	0	0.051
1	0.137	0.147
2	0.086	0.232
3	0.467	0.225
4	0.290	0.165
5	0.015	0.102
6	0.004	0.048
7	0.0003	0.021
8	0.0002	0.023
$9-10-11$	0	0.003

Table 4.2: The table represents frequency of nodes for each degree for the Road map network with 1800 nodes and an E-R network with the same number of nodes and same linking probability

4.2.2 Degree

Figure 4.10a represents the frequency of node for each degree and in tab. 4.2 is it possible to find the numerical values of the distribution compared with the ones of the E-R network.
The distribution of the nodes' degree is very peculiar especially compared to the E-R graph. Both the distribution are not flat but picked. The mode for the random distribution is 2 and it has a Gaussian bell shape. The 23.2% of the population have degree equal to 2 and the 39% of the population have degree equal to 3 or 4. The two tails are almost symmetric with the 19.8% of the population on the left and the 18.0% on the right. Instead the strength distribution of the road map is more asymmetric, more than the 75% of the nodes have degree equal to 3 or 4. On the right side the distribution is truncated,nodes with higher degree are almost impossible to find. The rest of the node 22% are on the left side. There are no isolated vertices and local peck is present for degree equal to one.
From the plot 4.11a it is possible to observe that nodes with degree in the range 1-4 are present in almost all the communities, while 97% of the ones with degree

$Betweeness$	$ratioRoad-map$	$ratioE-R$
$0.05 <$	0.001	0
$0.04 - 0.05$	0.002	0
$0.03 - 0.04$	0.003	0
$0.02 - 0.03$	0.004	0
$0.01 - 0.02$	0.014	0
$0.001 - 0.01$	0.098	0.098
$0.001 >$	0.837	0.902

Table 4.3: The table represents frequency of nodes with betweeness centrality in different range for the Road map network with 1800 nodes and an E-R network with the same number of nodes and same linking probability

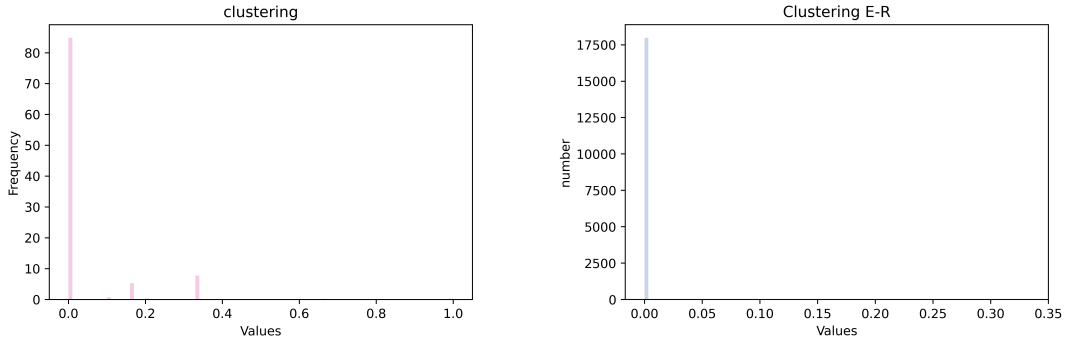
equal to 5 or 6 are in first half communities (the major ones). Node with degree equal to 7-8 are in the first six biggest communities. Just one node out of eight is in the 78-th community. Instead for the random network the communities have equally distributed the nodes'degree, with the exception of the 0 degree (for obvious reasons) and 1.

4.2.3 Betweeness centrality

The distibution of the betweeness centrality is pecked on 0 with a mean value of 0.001 ± 0.004. The distribution for the E-R is even more pecked the main value is $(3.8 \pm 4.7)\dot{1}0^{-4}$. It means that in the road map has more vertices with structural relevance than the E-R map in which is less important which node chose in order to shorten the path. In addition looking at 4.11c the first 25 five points with the highest betweeness belong to comunities 3,4,24. While in the E-R top 25 vertices belongs to 20 different. This fact underline that the betweeness relevance in the road is not only punctual but local.

4.2.4 Closeness centrality

In the E-R network the 94.4% of the population belong to the giant component. In this huge subset, due to the random structures, the position of the nodes makes no difference so in each community the the possible distance value are equally distributed. The main value for the giant component of the random graph is 0.104 ± 0.013. While for the real one, due to the absence of a true giant component the results related to the distribution of the closeness centrality are not so easy to understand. The distribution present two group of pecks. On the left the ones related to small components with mean value equal to 0.003 ± 0.003. On the right there is the peck related the biggest components with mean value 0.015 ± 0.002, the 64% of the population contributes to this peck. The node of this graph are more far from each others than the E-R's ones, even considering the biggest components. Indeed, it has a mean value equal to 0.023 ± 003 for the road network and 0.108 ± 0.008 for the E-R. What is even more different from the random network is the diversification



(a) *California road network* **(b)** *E-R network*

Figure 4.9: Clustering histograms of the Road map and E-R network

on values of the different communities 4.11e, also inside the same component. It means that the link structure varies for the different area of the network, not only because there are many not connected component but also because there is some topological issue.

4.2.5 Clustering

The random graph shows almost no clustering characteristic, in comparison the road map tend to create more tightly knit groups. The 7.2% of the population has a clustering coefficient equal to 0.167 while a 9.1% of the nodes has a value of 0.33. There are not any other relevance information related to the clustering measure.

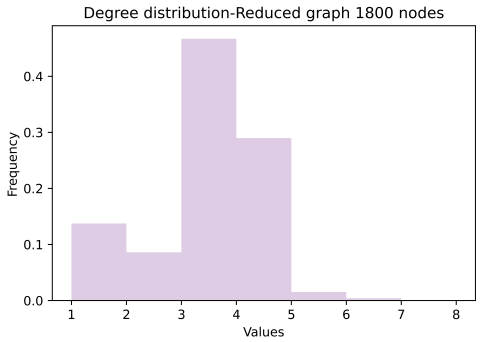
4.2.6 Further information

4.2.6.1 Betweeness vs degree

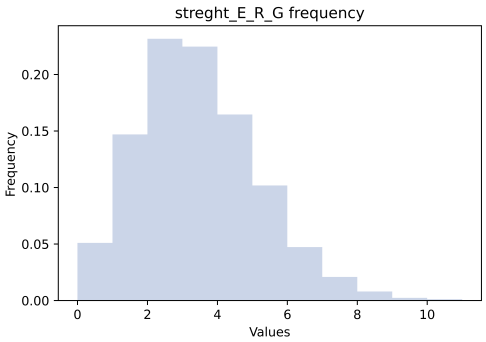
In the E-R there's a correlation between nodes degree and their betweeness the correlation coefficient is equal to 0.867. For the road map instead it is only 0.170. Tab 4.4 shows further information about the distribution of the betweeness for each degree. The distributions of betweeness for each degree are highly not symmetric. Furthermore as both the table 4.4 and the figure 4.12c the percentage of outliers (three σ far from the mean value) is very high, at least three times the percentage of outliers expected for a normal distribution (0.3 %)

4.2.6.2 Closeness vs Degree

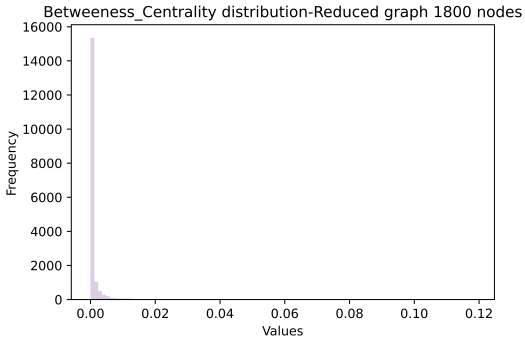
In the E-R there's a correlation between nodes' degree and their betweeness the correlation coefficient is equal to 0.596. For the road map instead it is only -0.153.



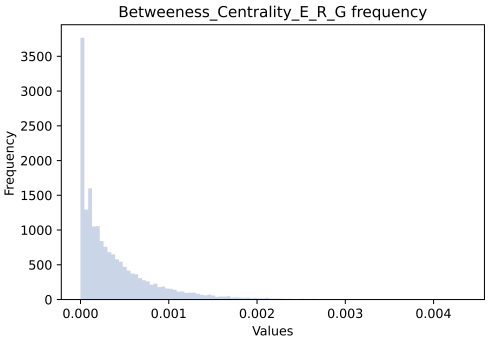
(a) *California road network*



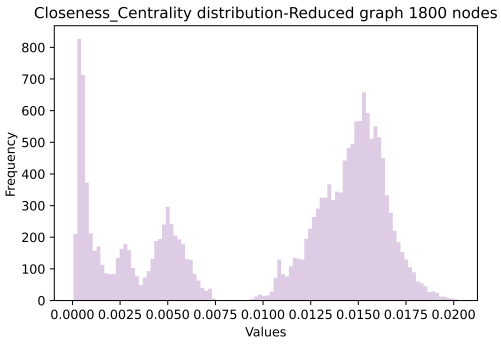
(b) *E-R network*



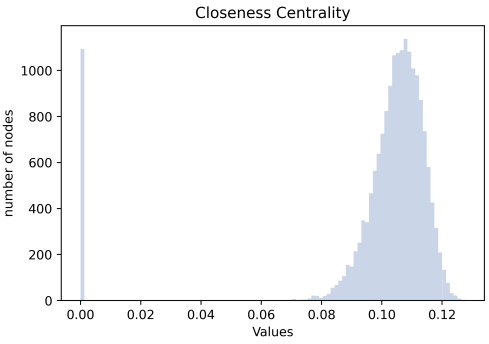
(c) *California road network*



(d) *E-R network*

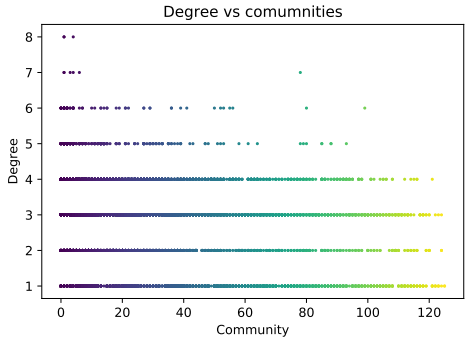


(e) *California road network*

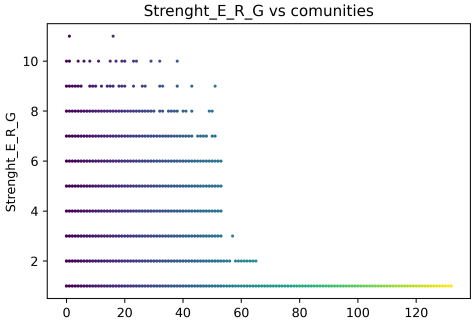


(f) *E-R network*

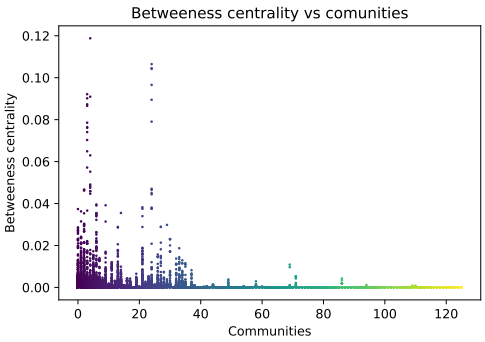
Figure 4.10: Degree, betweeness, closeness histograms of the Road map and E-R network



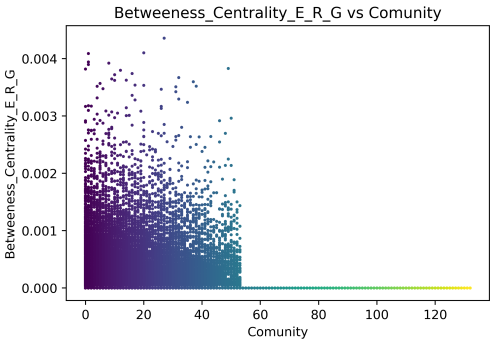
(a) *California road network*



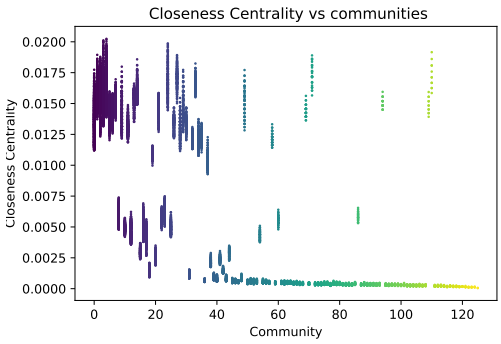
(b) *E-R network*



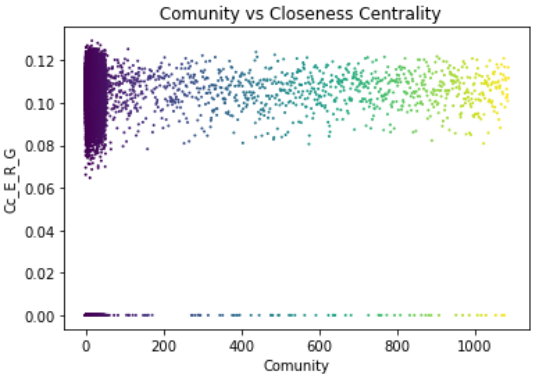
(c) *California road network*



(d) *E-R network*



(e) *California road network*



(f) *E-R network*

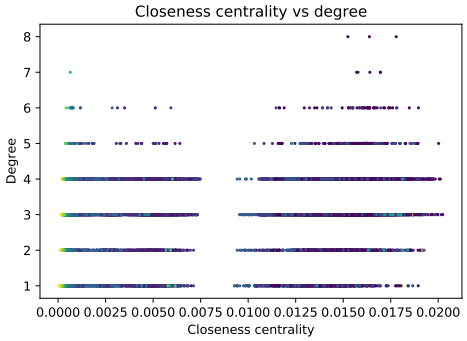
Figure 4.11: Degree, betweeness, closeness scatter plots of the Road map and E-R network

$Degree$	$Mean - betweeness$ 10^{-3}	σ 10^{-3}	max 10^{-3}	$outliers$ %
0	0	0	0	0
1	0	0	0	0
2	0.2	1.4	28.8	1.7
3	0.9	4.1	118.8	0.9
4	1.82	5.17	104.5	1.6
5	6.30	8.07	62.3	2.6
6	4.29	5.32	21.3	1.4
7	7.15	1.14	32.2	0
8	5.77	3.46	10.4	0

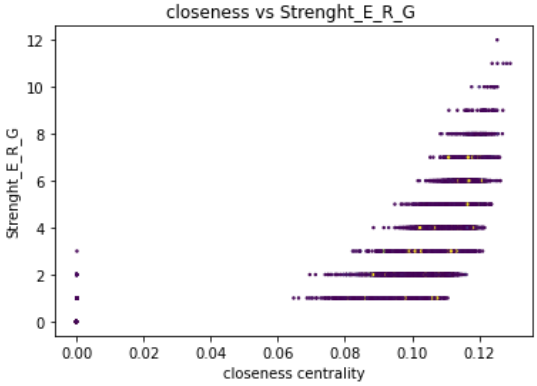
Table 4.4: California road map betweeness vs degree. The column three represents the standard deviation of the betweeness of the sample, the fourth is max value of betweeness for each degree, while the last one is the percentage of nodes with a value of it three times the standard deviation away from the mean value

$Degree$	$Meanbetweeness$ (10^{-3})	σ (10^{-3})	max (10^{-3})	$outliers$ %
0	0	0	0	0
1	0	0	0	0
2	0.11	0.06	0.46	0.37
3	0.3	0.1	1.2	0.26
4	0.5	0.2	1.6	0.17
5	0.8	0.3	20.6	0.08
6	12.1	0.4	31.8	0.04
7	16.4	0.5	35.1	0.02
8	21.5	0.7	38.3	0
9	25.9	0.7	43.6	0
10	30.5	0.7	41.0	0
11	31.1	0.6	37.4	0

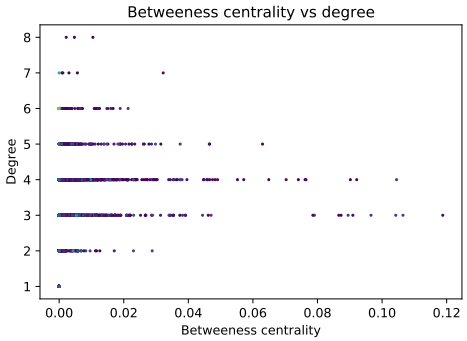
Table 4.5: E-R betweeness vs degree. The column three represents the standard deviation of the betweeness of the sample, the fourth is max value of betweeness for each degree, while the last one is the percentage of nodes with a value of it three times the standard deviation away from the mean value



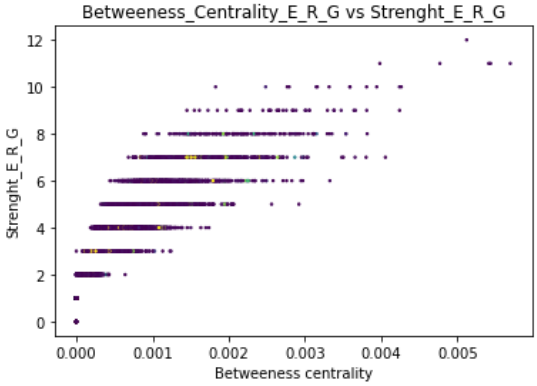
(a) *California road network*



(b) *E-R network*



(c) *California road network*



(d) *E-R network*

Figure 4.12: Centrality measures in function of degree for the Road map and E-R network

	$\Delta Road - map$	$\Delta E - R$
Betweenesscentrality	0.94%	0.413
Clustering	−0.90%	0
Degree	−0.24%	−0.44
$Shortest - path$	1.02%	0.70

Table 4.6: Percentage variation removing 11 nodes with the highest degree

	$\Delta Road - map$	$\Delta E - R$
Betweenesscentrality	17.8%	0.12%
Clustering	−0.03%	0.07%
Degree	−0.14%	−0.06%
Shortestpath	19.1%	0.11%

Table 4.7: Percentage variation removing 11 nodes with the highest Betweeness

4.2.7 Perturbation

4.2.7.1 Degree based

Table 4.6 shows the variation of average values removing 11 (0.06% of the population) nodes with highest degree:
The removal of nodes with highest degree has a major effect not on the average degree it self but on the others quantity. The perturbation had a percentage effect bigger than the percentage variation of the number of the removed nodes. The degree variation is almost the double for the E-R graph because the maximum degrees are higher then the ones of the road map. With the exception of the clustering, all the variation are of the same order.

4.2.7.2 Betweeness based

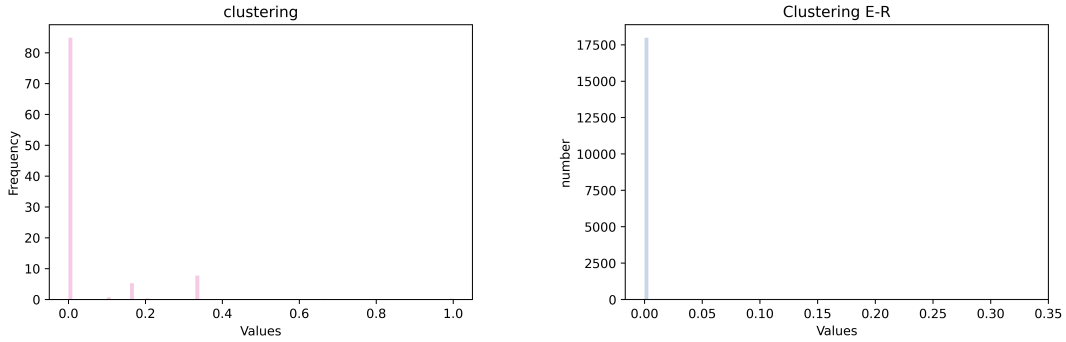
Table 4.7 shows the variation of average values removing 11 (0.06% of the population) nodes with highest betweeness:
The removal of nodes with highest betweeness has a major effect on the average betweeness and on the average shortest path and on these two quantities the perturbation had a percentage effect of around 300 times bigger than the percentage variation of the number of the removed nodes. This effect is totally different for the E-R network and so it underline the structural properties of these nodes.

4.2.7.3 Random

Table 4.7 shows the variation of average values removing 11 (0.06% of the population) nodes with highest betweeness.
In table 4.8 it is possible to see that the perturbation for the betweeness as a major effect on the E-R, this underline again the fact that in the road map fewer point has

s	$\Delta Californiamap$	$\Delta E-R$
$Betweenesscentrality$	0.19%	1.2%
$Clustering$	−0.2%	0.07%
$Degree$	−0.08%	−0.04%
$Shortestpath$	0.21%	0.90%

Table 4.8: Percentage variation removing 11 random nodes



(a) *California road network* **(b)** *E-R network*

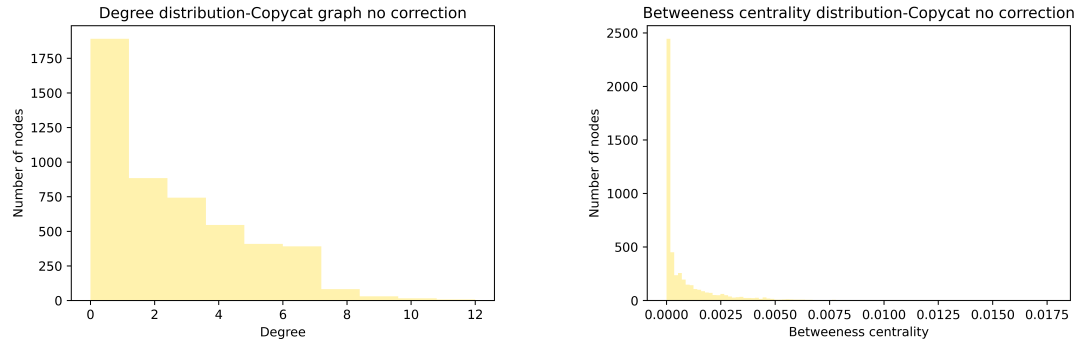
Figure 4.13: Clustering histograms for the road map and the E-R network

a major role and so it is more unlikely to change the mean value, while in a random network vertices have close values.

4.3 Copycat

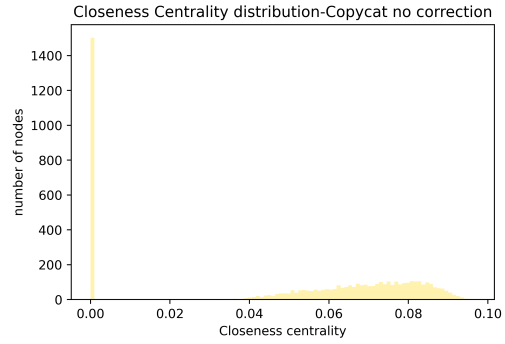
4.3.1 First steps

The distributions after the first step in the building of the copycat network are shown in 4.14. The network is full of isolated nodes, the number of each column decreases with the increasing of the degree. This distribution and the closeness one compared with a real graph of 5000 nodes has χ^2 test equal respectively of 14313 and 1125982, and their p-value are less then 0.01, this leads to the rejection of the hypotheses of compatibility of these distributions with the ones of the real network. Instead the betweeness distribution has a $\chi^2 = 191$, so even if it is > 23.6. So the hypotheses of compatibility has to be rejected, but his value is much lower than the other two. Indeed, also qualitatively the two seems more similar. Finally for the clustering distribution the $\chi^2 = 521$ with a p-value < 0.01. So the two distribution are not compatible. in conclusion, the first part of the algorithm fits better the betweeness features than the other ones.
The following results concerned the copycat network after all building process

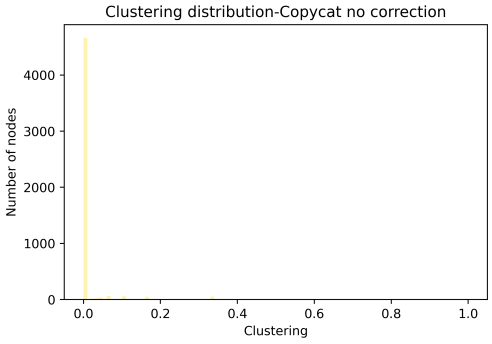


(a)

(b)



(c)



(d)

Figure 4.14: Main features histograms of the Copycat histogram at first ste

4.3.2 Degree

The two degree distributions (4.15a, 4.15b) are qualitatively similar but have a χ^2 test value of 29 and a correspondent p-value < 0.01, so the two distribution are not quantitatively compatible. A glance to the fig 4.16c and 4.16d shows the different behaviour of the degree for different communities. Indeed,in the road map also nodes in small communities have high degree nodes while in the copycat graph small communities have only nodes with degree equal to one. So even if the total distribution is enough similar with the empirical one the community structures underline a more complex network compared with the silicon one.

4.3.3 Betweeness centrality

The two betweeness distributions (4.15c) have a χ^2 test value of 1189 and a correspondent p-value of < 0.01, so the two distribution are not compatible. Comparing this value with the on obtained in the first step of the algorithm is it possible to state that during the process of bonding isolated nodes the betweeness distribution get even farther from the real distribution. There are more differences in the distributions to underline. In the real one the ratio of outliers is $(2.1 \pm 0.4)\%$ (the value is calculated from 15 different subset of the real network) is more than the double of copycat one 1.15% and therefore not compatible. Furthermore looking at the figures 4.16e and 4.16f it is possible to assert that qualitatively the copycat network is more uniform than the real one.Indeed taking in consideration separately each community the mean value of the standard deviation of the betweeness of the first 20 communities is the same, but the standard deviation of the standard deviation of the real graph is 0.0016 while the other one is 0.0003. So the the spread of the data of different communities is more different in the empirical network.

4.3.4 Closeness centrality

Also the two closeness distributions (4.16a) have a p-value < 0.01 with a χ^2 test value of 2092 and so the two distribution are not compatible. It was almost impossible to model the peculiar shape of the closeness centrality scatter plot of the nodes divided in communities as shown in 4.11e. This characteristic as explained in 4.2.4 is related to the topological structures of each communities. The model was not able to copy this feature even qualitatively.

4.3.5 Clustering

The two clustering distributions (4.15a) have a χ^2 test value of 442 and a correspondent p-value < 0.01, so the two distribution are not compatible. But it behaves in a more similar way to the real network than a E-R network, indeed a χ^2 test for the random and copy graph shows a value of 4532056. The main difference of the road networks and the copycat network is related to distribution of clustering among communities. Indeed,in the real case also small communities have nodes with clustering coefficient > 0 while in the copycat network this not occurs. This underline a district characteristic typical of a road map where also points belonging to small

Degree	$BCRoad-map$	$Road-mapCopycat$ 10^{-3}	*compatibility*
0	0	0	0
1	0 ± 0	0.034 ± 0.14	0.0534789
2	0.34 ± 1.3	1.04 ± 0.8	0.207949
3	0.8 ± 2.1	2.6 ± 1.6	0.418264
4	1.7 ± 3.6	4.5 ± 2.1	0.459305
5	1.0 ± 1.9	4.4 ± 2.6	1.13375
6	1.1 ± 1.8	13.2 ± 3.4	9.77051

Table 4.9: In column two and three are represented the mean values of the betweeness centrality (BC)

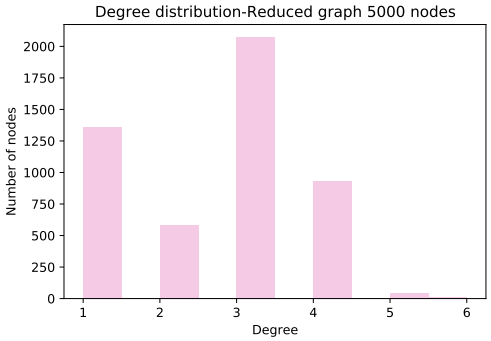
communities tends to create connection among nodes with common neighbors. This not happens in the copycat model.

4.3.6 Betweeness vs degree

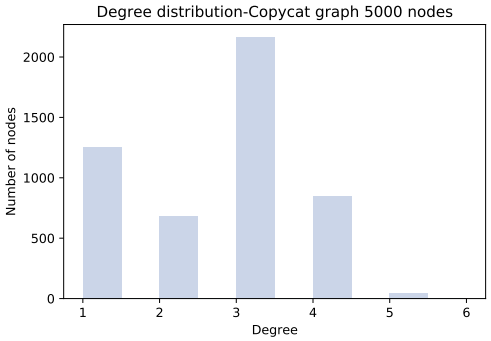
Even if the betweeness distributions are not compatible, the scatter plot of the betweeness vs the degree is an interesting tool in order to see how much this network construction is qualitative similar to real one or has some features similar to the E-R. For the copycat points of the plot 4.18b there is a correlation of 0.73 while for the California one is 0.25. So the copycat nodes with high degree have a more probability to have high degree like in a E-R graph, while for the real network the betweeness is not related to the degree. The idea to compare the distribution of betweeness of nodes with the same degree by comparing his mean value and standard deviation leads to a blind spot. Table 4.9 shows how wide is the spread of the points around the mean, the standard deviation are so big to make almost all mean values compatible (the compatibility was calculated with the square of the differences of mean valued divided by the sum of the variances). Table 4.10 shows the number of the outliers of the betweeness distribution of node with the same degree. The high number of them underline a non random feature in the construction of copycat model. Indeed it has shown that for E-R network the percentage is lower. Anyway the number of outliers for each degree is not compatible with the real network. The copycat graph needs more node with degree equal to 3 and high betweeness and less node with degree node equal to one and high betweeness.

4.3.7 Closeness centrality vs degree

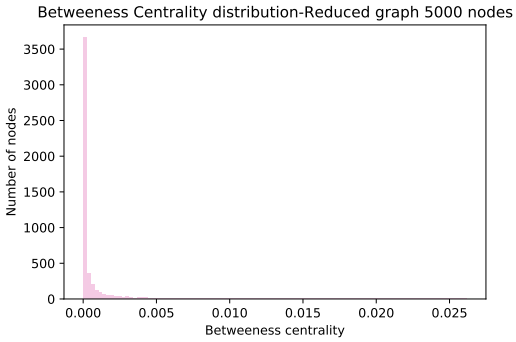
For completeness the plot of closeness vs degree is shown in 4.18d. The real network has a correlation coefficient of 0.21 while the copycat value is 0.48. In the California map till degree 4 the centrality values of the nodes take almost all the possible value of the whole distribution (4.18c), instead for the model, due the light correlation nodes with high degree tends to have high closeness and vice-versa



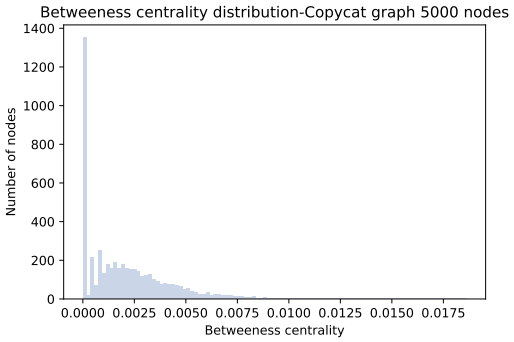
(a) *Road map*



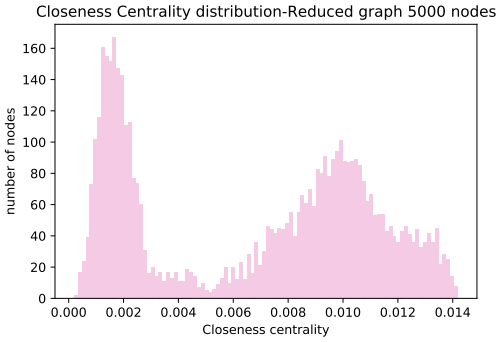
(b) *Copycat network*



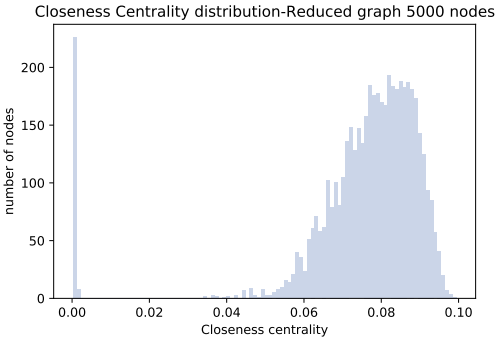
(c) *Road map*



(d) *Copycat network*

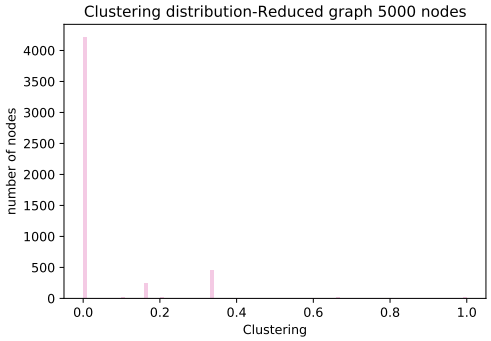


(e) *Road map*

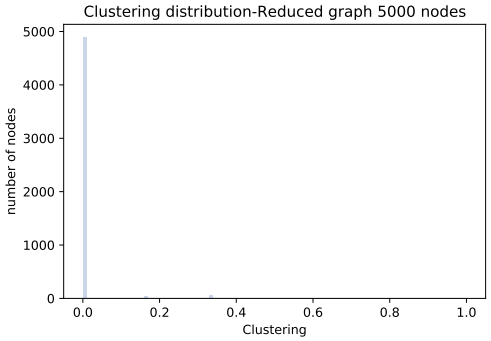


(f) *Copycat network*

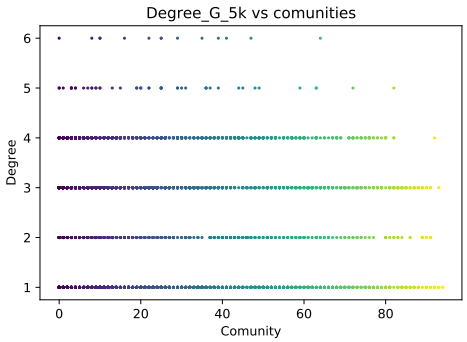
Figure 4.15: Degree, betweeness, closeness histograms for the road map(5K nodes) and the E-R network



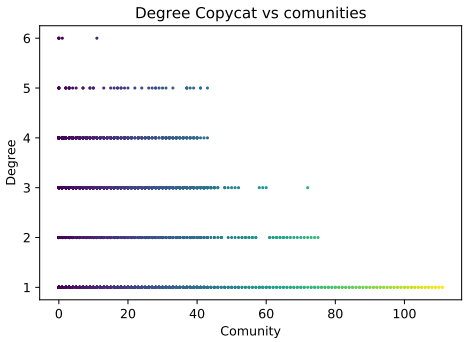
(a) *Road map*



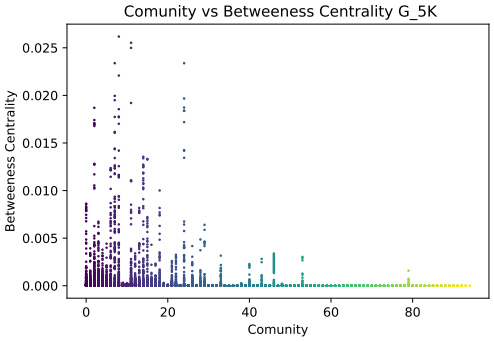
(b) *Copycat network*



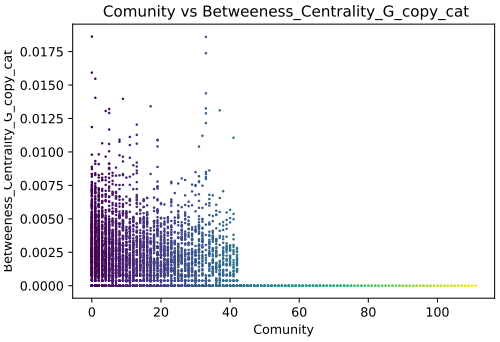
(c) *Road map*



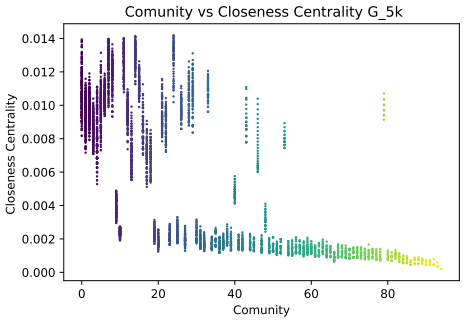
(d) *Copycat network*



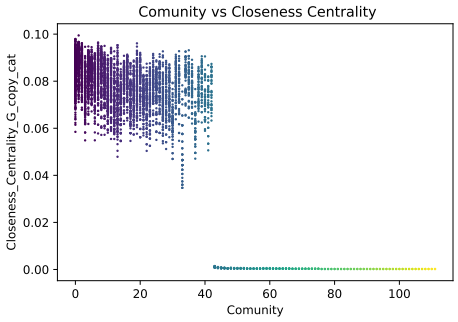
(e) *Road map*



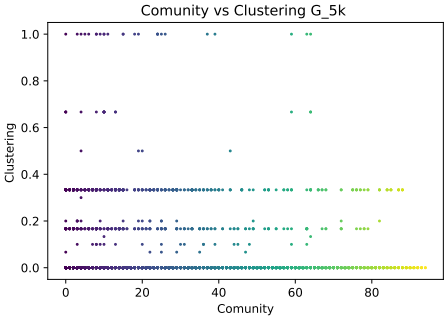
(f) *Copycat network*



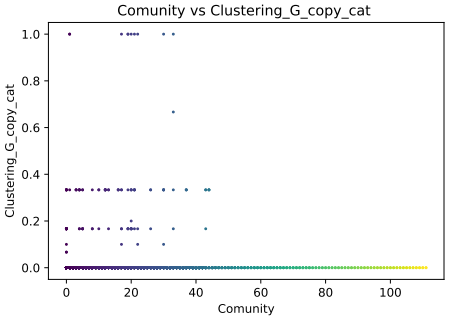
(a) *Road map*



(b) *Copycat network*



(c) *Road map*



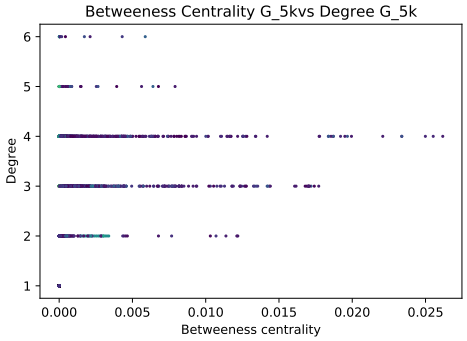
(d) *Copycat network*

Figure 4.16: Closeness and clustering scatter plot for the road map(5K nodes) and the Copycat network

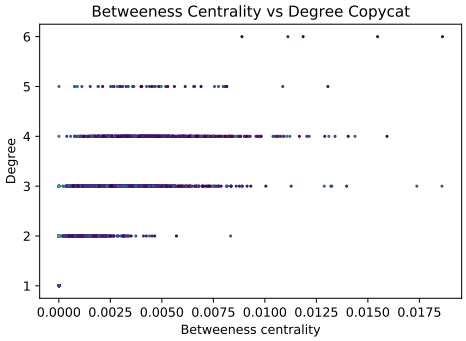
Figure 4.17: Clustering histograms and degree and betweeness scatter plot for the road map(5K nodes) and the E-R network

Degree	$OLRoad-map$	*OLCopycat*
0	0	0
1	0	0.8
2	0.2	0.16
3	1.3	0.6
4	0.5	0.2
5	0.04	0.02
6	0	0

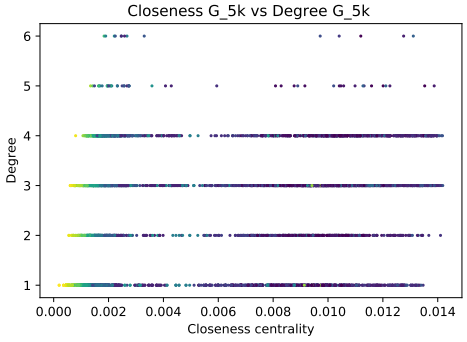
Table 4.10: In columns two and three are represented the frequency of outliers for each degree, nodes whose betweeness is more than three standard deviation away from the mean value



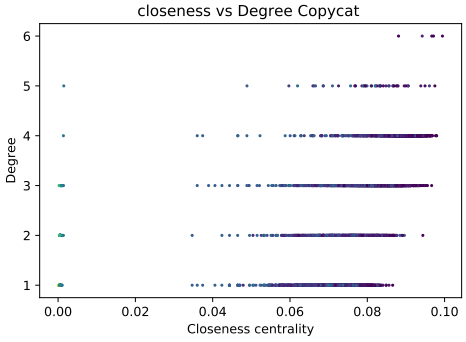
(a) *Road map*



(b) *Copycat network*



(c) *Road map*



(d) *Copycat network*

Figure 4.18: Centrality measures vs degree for the road map(5K nodes) and the E-R network

5

Conclusion

5.1 Degree

From 7236 nodes the ranking of the number of nodes with the same degree is invariant and form 1.810^4 nodes the average variation of the ratio of each degree respected to the final value is equal to 0.8%. Considering the χ^2 test the compatibility is reached after size=1'788'847.. The distribution both for the whole network both for the reduced one (1800 nodes) is not gaussian, unlike a E-R whose distribution is normal and pecked on the value of 2, but it is asymmetric. More then the 75 % of the population has a degree of 3 or 4 with practically no tail after the value of 5. This leads to the conclusion that there is some kind of rule in the building of the road which makes very unpractical the linking of a road with more then 4 roads.

5.2 Betweness centrality

Qualitatively this features is size invariant,for at list network whose nodes are in the range 2139-27449, so the behaviour of the cumulative distribution is the same with compatible mean values. Quantitatively, in this range, the $\chi^2 = 144 \pm 121$. With regard to the reduced graph, the shape of the distribution is similar to the one of a E-R graph with the peculiarity of a long tail, a bigger numbers of outliers and a mean value almost three times bigger than the E-R's one. Most of the nodes, 84% of the population, has a betweeness less than 0.001. The correlation between degree and betweeness is 0.17. In opposition with a random graph, nodes with high betweeness are not the ones with high degree. Another different is related to the fact that all the first 25 nodes with the highest value belongs to communities 3,4 and 24; in the random network the first 25 are spread on 20 different communities. Furthermore a small number of vertices, 0.06% of the total, has a major role, indeed by removing them the average values of betweeness and shortest path increase of 17.8% and 19.1%. The high number of outliers, the low correlation with the degree and their role in the networks leads to the conclusion that the linking of this nodes is not something random but related to some inner information of the graph it self (bridges, highways, tunnels)

5.3 Closeness centrality

The distribution of this value changes a lot with the increasing of the size, one of the main reason is the absence of a real giant component. So this feature is neither qualitatively nor quantitatively size invariant. the closeness values are not just related to the components of which the nodes belongs to, but also different communities of the same component can have slightly different range of values. This is in stark contrast with a random network in which all the communities have the same range of values. This underline some kind of issue related to the topological structure its self, which does not depend only by the euclidean distance. Furthermore the correlation of the closeness with the degree has Pearson coefficient of -0.15 while for the E-R it is equal 0.60. So the centrality of the node it is not strictly related to his degree as in the case of a random network.

5.4 Clustering

The clustering distribution is qualitatively invariant under size transformation, the ranking of vertices with the same clustering coefficient does not change and the the proportions of nodes with different value are similar. Quantitatively the χ^2 value among distributions of different size (in the range $1.7 \cdot 10^4$ - $1.9 \cdot 10^6$) is 247 ± 307. A χ^2 test compatibility occurs for size bigger than 1'567'683 with a value equal to $19 \pm 13 < 23.6$. Comparing the reduced graph of the road map (1800 nodes) with a E-R map with the same size and same linking probability the first tends to create more tightly knit groups. With respect to the reduced network, the 7.2% of the population has a clustering coefficient equal to 0.167 while a 9.1% of the nodes has a value of 0.33. Instead the E-R graph shows almost no clustering characteristic, 99.97% of the nodes has a clustering value of 0. Finally, as expected among the three kind of perturbation (degree based, betweeness based and random) the one which has a bigger effects is the degree based. Removing the 0.06% of the sample with the highest degree a variation of -0.90% occurs.

5.5 The copycat network

The linking probability between two nodes in function of their distance (previously modeled by the $spring_layout$ function of the packages Networkx) allowed to rebuild a copy of the road map. This first step achieved a qualitative similarity for the betweeness distribution, but not quantitatively the $\chi^2 = 191 > 23.6$. The other features was not even qualitatively near to the real one. A possible explanation is that the linking probability is based only on the distance, so it allows to fit the right number of nodes which are fundamental in the construction of the shortest paths. On the other hand it not takes into account a probability distribution of the number of edges for each node and so could happened that most of them has 0 links or a too much high degree in relation of what it is expected. Indeed, the previously analysis has shown that betweeness and closeness are correlated with the degree in the case of random graph. In the road network nodes in the middle of the map, so

surrounded by lots of nodes, have most of the time the same probability to have the same degree of nodes set near the periphery. In this model case node in the middle have a higher number of try to make a connection.
The number of degree has a consequences on the closeness, if vertices are allowed to be isolated the will not be able to reached or been reached by the others one. The incompatibility of the clustering distribution can be explained by the randomness of the process. In a road map it easily imaginable the presence of districts, instead the process of this algorithm does not take into account this aspect and so it follows a clustering distribution similar to the one of an E-R graph.

The second step of the process almost reached easily the degree distribution compatibility $chi^2 = 29 > 18.5$ (where 18.5 is the value under which the hypothesis of compatibility for χ^2 with 6 degree of freedom can be accepted . However, this process has taken the the betweeness distribution away form the real one. indeed, the number of nodes with betweeness > 0 increases. It is hard to explain why and what should be the expected distribution. The algorithm in a not quantified way cut links or add new ones with a preferential attachment rule related on the distance probability function when it is possible, otherwise the linking will be random. This leads to a greater number of them acting as bridges among them, joying far vertices. More analysis should be done on it in order to understand which part of the process has which consequences on the final betweeness distribution.

Due to the connecting of the isolated nodes and the increase of the ratio of degree three and four, the number of nodes with closeness centrality > 0 increases. However the the distribution reached is not compatible with the real one even qualitatively due to the lack of the two pecks in the distribution. Also here the explanations could be several. The starting position of the node using the *spring_layout* function is not suitable and so from that position no rule could not be done in order to join the points correctly. Alternatively the positions are well set but he preferential attachment rule chosen is not enough correct. Finally the incompatibility could be caused by all the exceptions in which the rule could not be follow and random nodes with a selected degree were chosen to be joined. However as it was shown the closeness centrality among the different communities shows a very peculiar behaviour, difficult to copy so with no doubt a more fine work is needed in order to reach some appreciable result.

Finally the clustering features shows a behaviour more similar to the road map than an E-R network. However quantitatively the real and model graph are not compatible. Qualitatively the Copycat network has clustering coefficient equal to 0 for nodes that belong to small communities in opposition with the road map where a district characteristic occurs also in small communities.