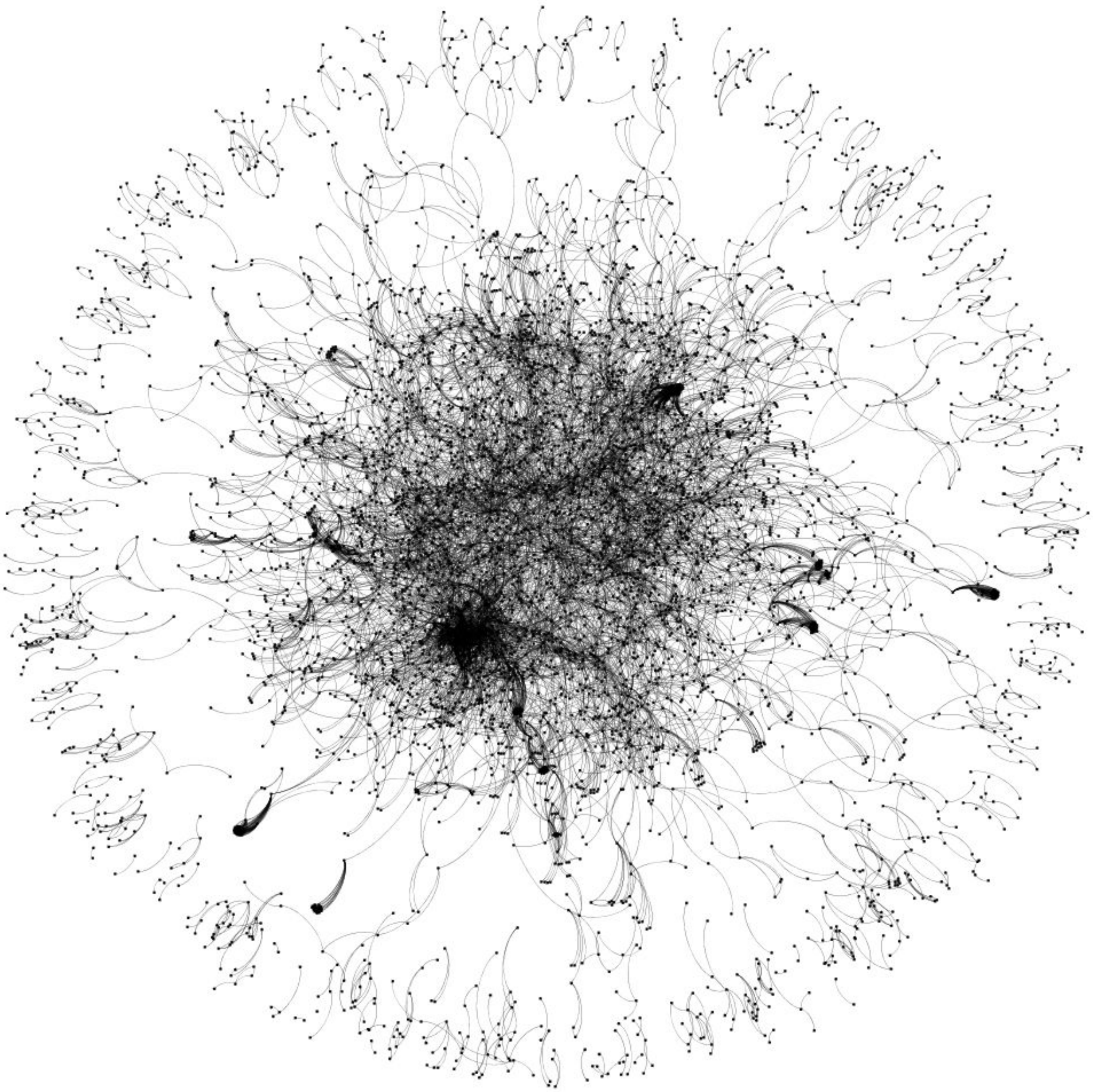
Graph Analytics class, a.y. 17/18

Second Assignment

Analyse your “realistic” graph

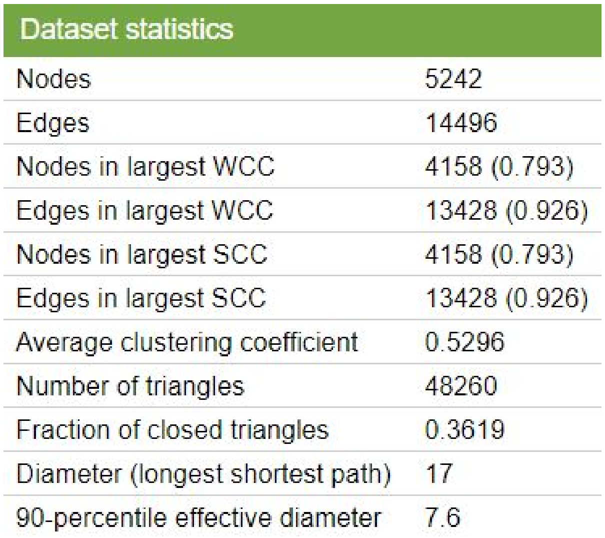


Cantarini Giorgio [s3828113]

Franco Danilo [s3809721]

0. Dataset Information

Arxiv GR-QC (General Relativity and Quantum Cosmology) collaboration network is from the e-print arXiv and covers scientific collaborations between authors papers submitted to General Relativity and Quantum Cosmology category. If an author i co-authored a paper with author j, the graph contains an undirected edge from i to j. If the paper is co-authored by k authors this generates a completely connected (sub)graph on k nodes.



Dataset statistics	
Nodes	5242
Edges	14496
Nodes in largest WCC	4158 (0.793)
Edges in largest WCC	13428 (0.926)
Nodes in largest SCC	4158 (0.793)
Edges in largest SCC	13428 (0.926)
Average clustering coefficient	0.5296
Number of triangles	48260
Fraction of closed triangles	0.3619
Diameter (longest shortest path)	17
90-percentile effective diameter	7.6

The data covers papers in the period from January 1993 to April 2003 (124 months). It begins within a few months of the inception of the arXiv and thus represents essentially the complete history of its GR-QC section.

1. Node level measurements

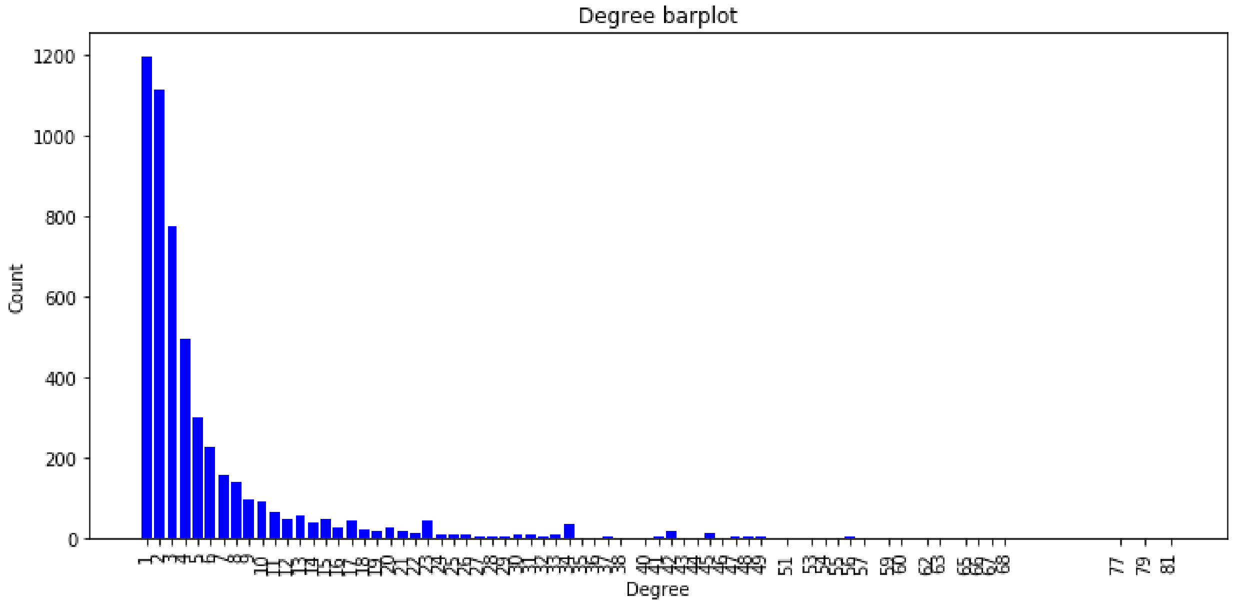
1.1. Degree distribution

Provides the probability that a randomly selected node in the network has degree k:

$$p_k = \frac{N_k}{N}$$

Degree distribution:

- average: 5.53
- variance: 62.70
- maximum: ('21012', 81)
- minimum: ('24372', 1)
- median: ('19454', 3)



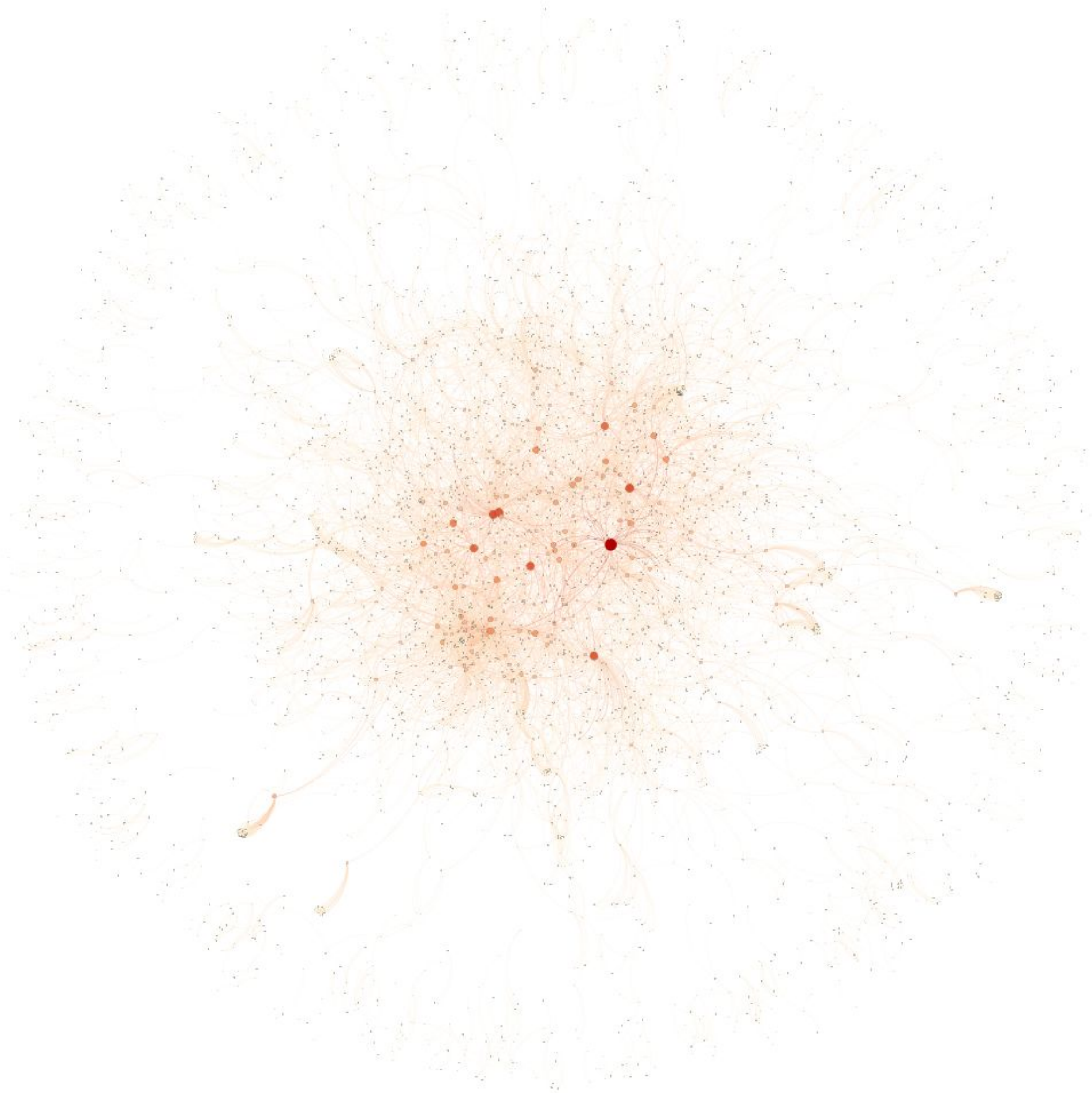
Looking at this histogram, we could observe that a fraction of nodes has a very high degree; the presence of hubs that have much larger connectivity than the average is a characteristic of power-law networks.

In fact, for a random network, we can state that the maximum degree node should yield:

$k_{max} \approx \langle k \rangle + \sqrt{\langle k \rangle}$ [mean + stdev for a Poisson distribution, 5+2.23 in the example],

but we can observe many nodes with degree greater than 10 (646 out of 54242 exactly) with a peak of 81 connections; moreover, we can also recognize the distinctive power-law distribution.

1.2. Betweenness



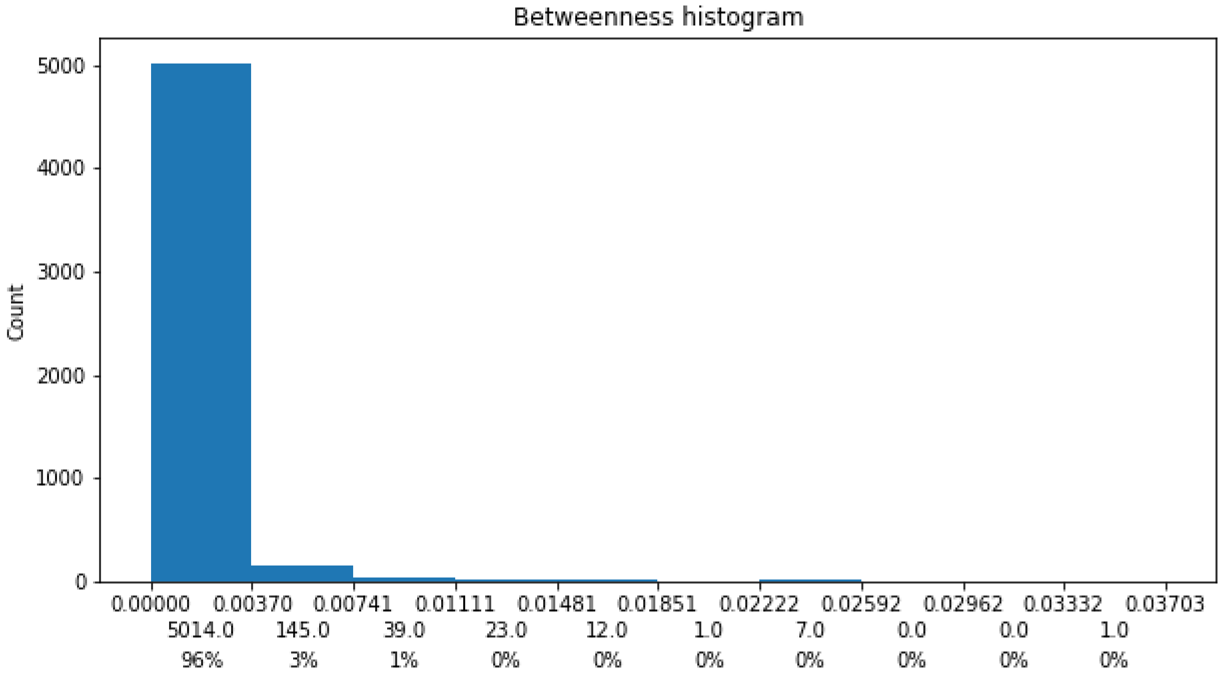
Measures the extent to which a vertex lies on a path between other vertices:

$$B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}},$$

where σ_{st} is the total number of shortest paths from any node s to any node t in the network and $\sigma_{st}(v)$ is the number of those paths that pass through v.

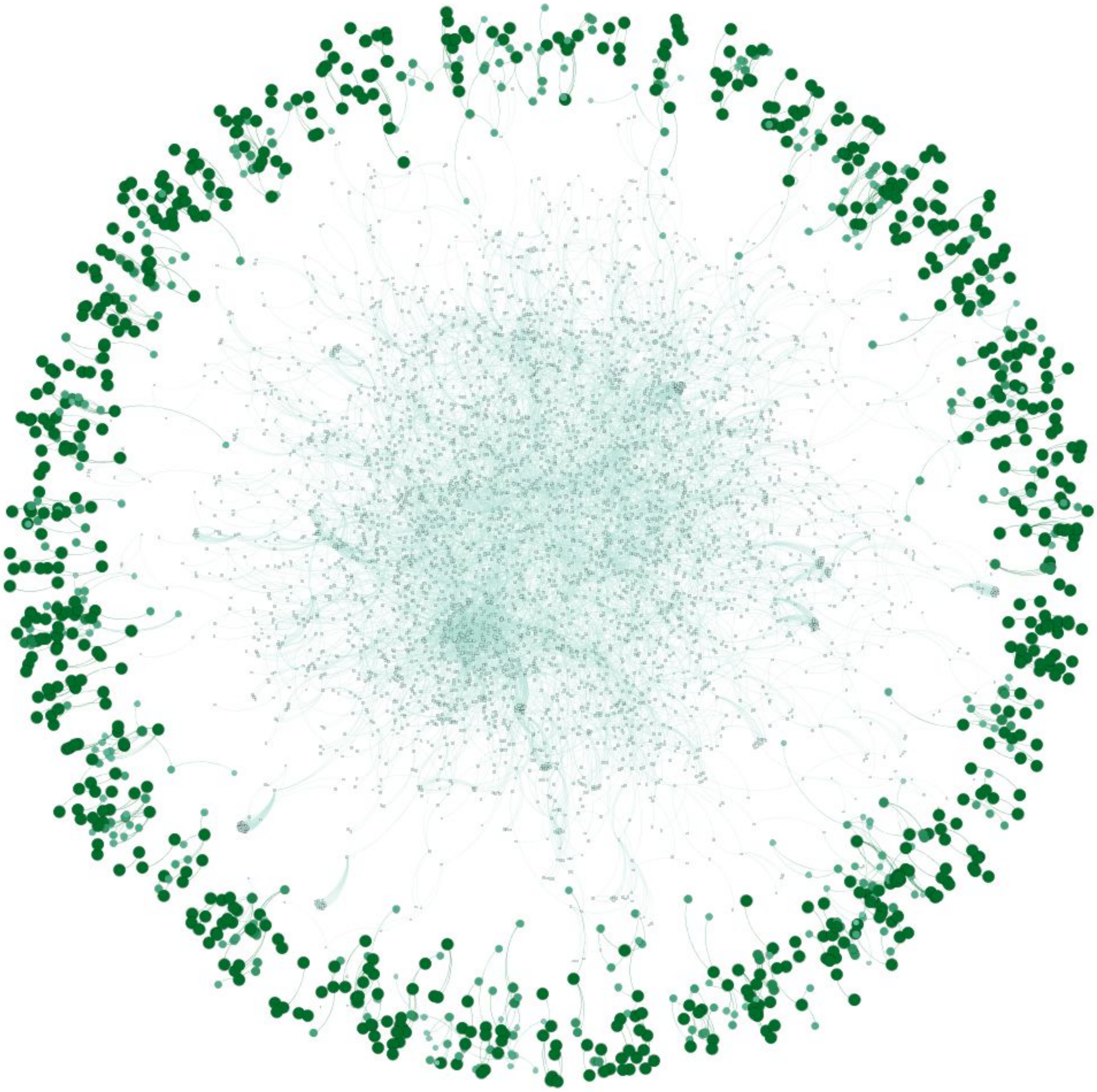
Betweenness distribution:

- average: 0.00
- variance: 3.89e-06
- maximum: ('13801', 0.04)
- minimum: ('5233', 0.0)
- median: ('21594', 0.0)



There are few nodes with relatively high betweenness which have a considerable influence in a network (their removal may disrupt communication); this suggests that the network is scale-free with few hubs from which the majority of communication pass through.

1.3. Closeness



Measure the mean distance of a vertex to other vertices:

$$C(x) = \frac{1}{\sum_y d(y,x)},$$

where $d(y,x)$ is the geodesic distance between vertices x and y.

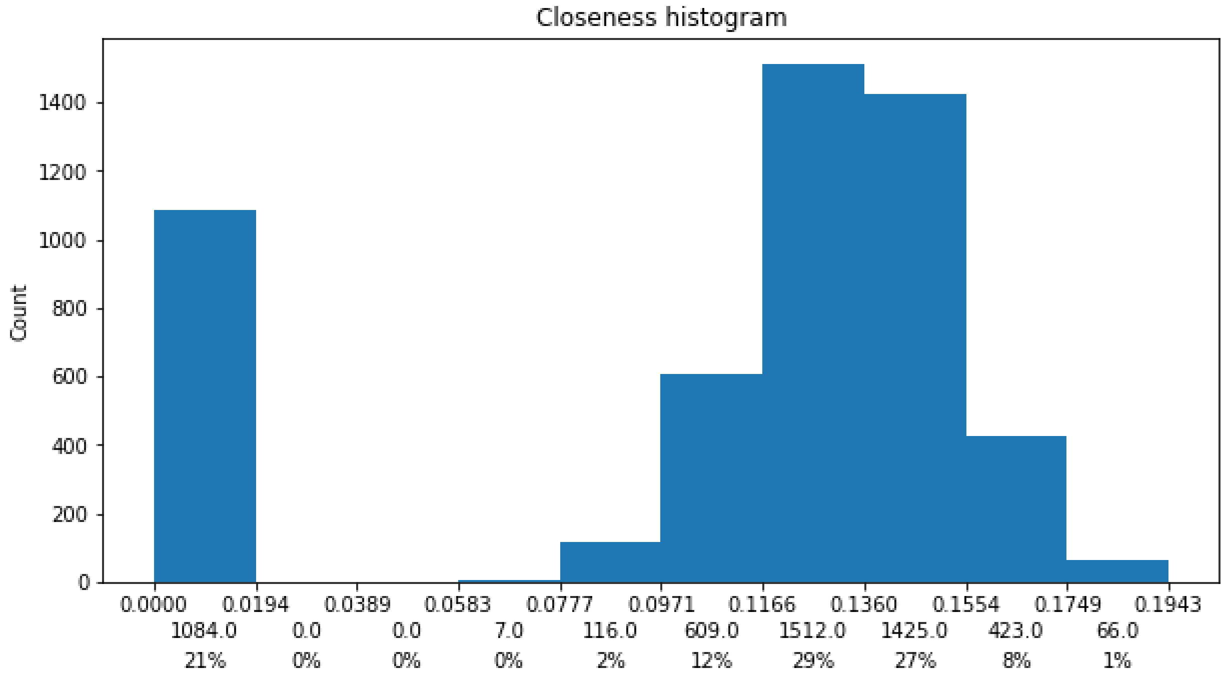
N.B.: In the case of a graph with more than one connected component, the NetworkX library does not use the closeness centrality as defined above, but compute a variant of it (*Wasserman and Faust improvement*):

$$C_{WF}(x) = \frac{n-1}{N-1} \cdot \frac{n-1}{\sum_y^{n-1} d(y,x)},$$

where n is the number of nodes reachable from x, while N is the total number of nodes (in fact this version compute a size-scaled centrality measure).

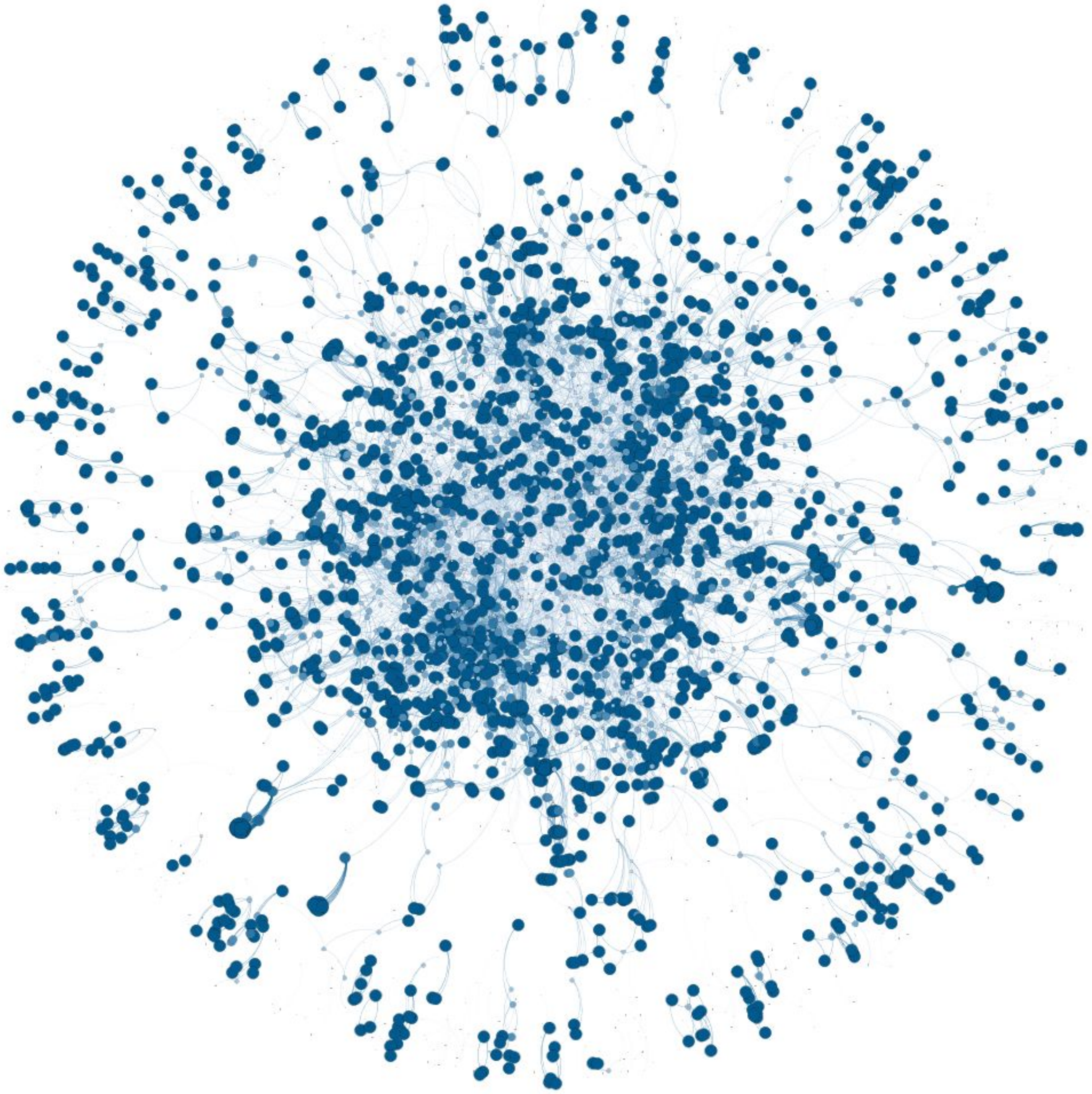
Closeness distribution:

- average: 0.11
- variance: 0.01
- maximum: ('13801', 0.19)
- minimum: ('12295', 0.0)
- median: ('2607', 0.13)



Nodes with high closeness have more direct influence on the other vertices; from this point of view we can easily tell the giant component apart from the periphery thanks to the Wasserman and Faust weighted centrality: nodes belonging to the connected component will have more influence on the graph dynamics (thus reflecting in a higher closeness) respect to the isolated clusters.

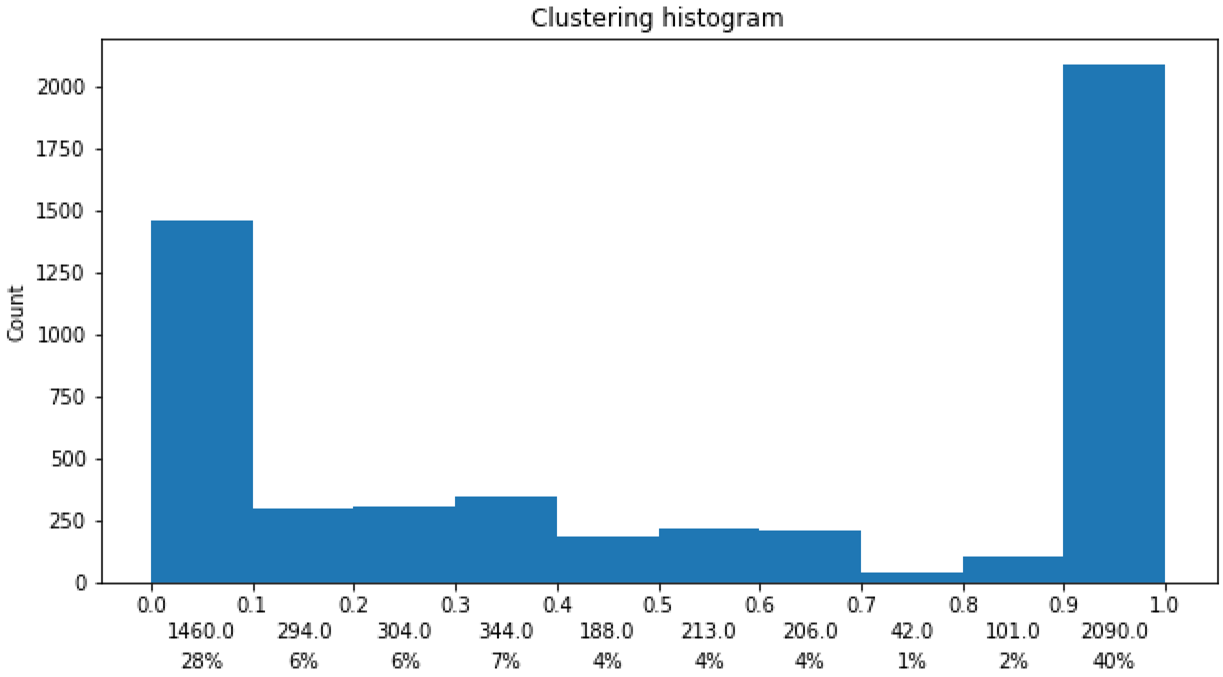
1.4. Clustering



The coefficient that captures the density of links in node i's immediate neighbourhood.

Clustering distribution:

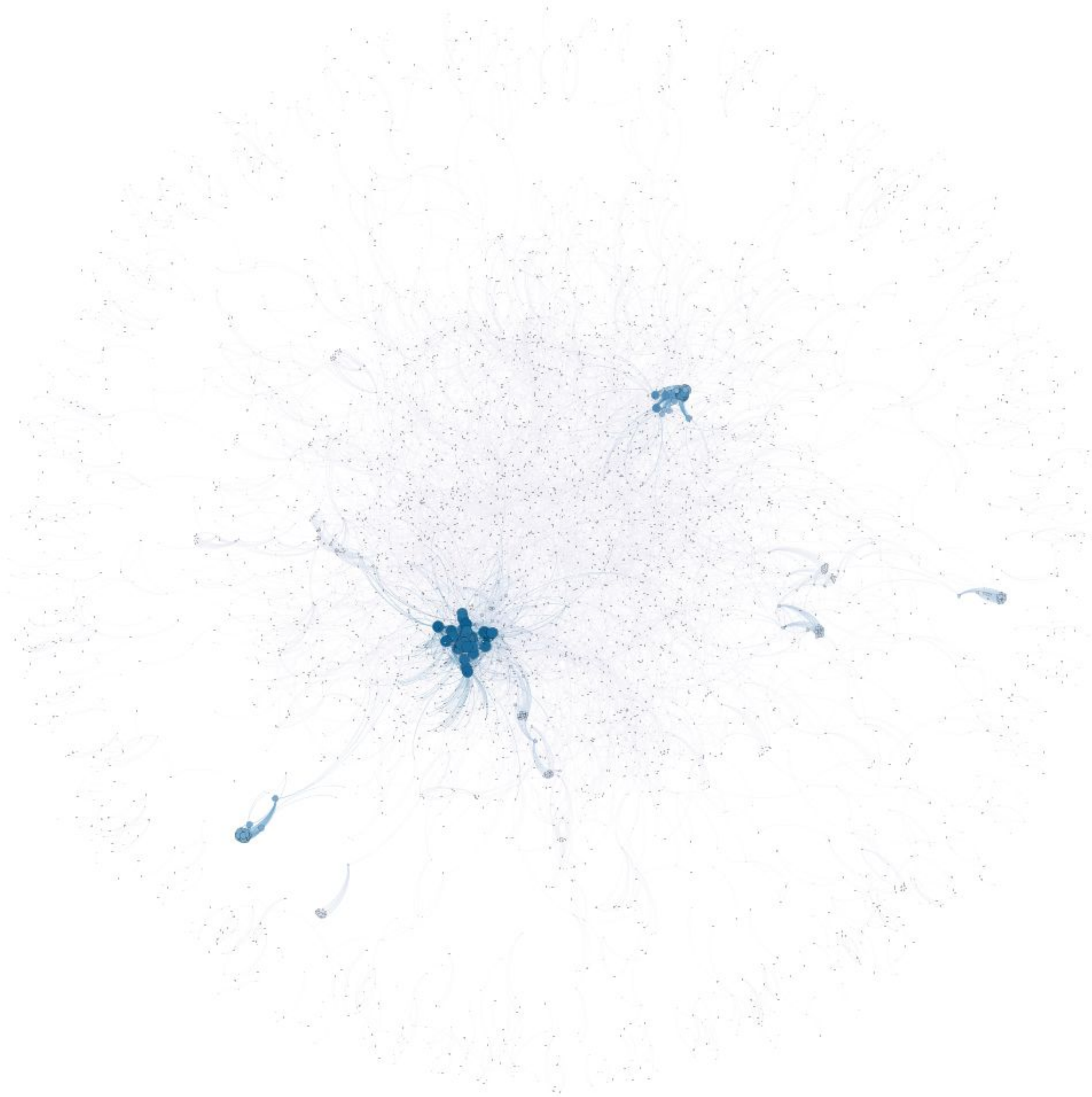
- average: 0.53
- variance: 0.18
- maximum: ('19521', 1.0)
- minimum: ('24372', 0)
- median: ('6971', 0.5)



Local clustering can be seen as an indicator of structural holes in the network: small values indicate powerful individuals.

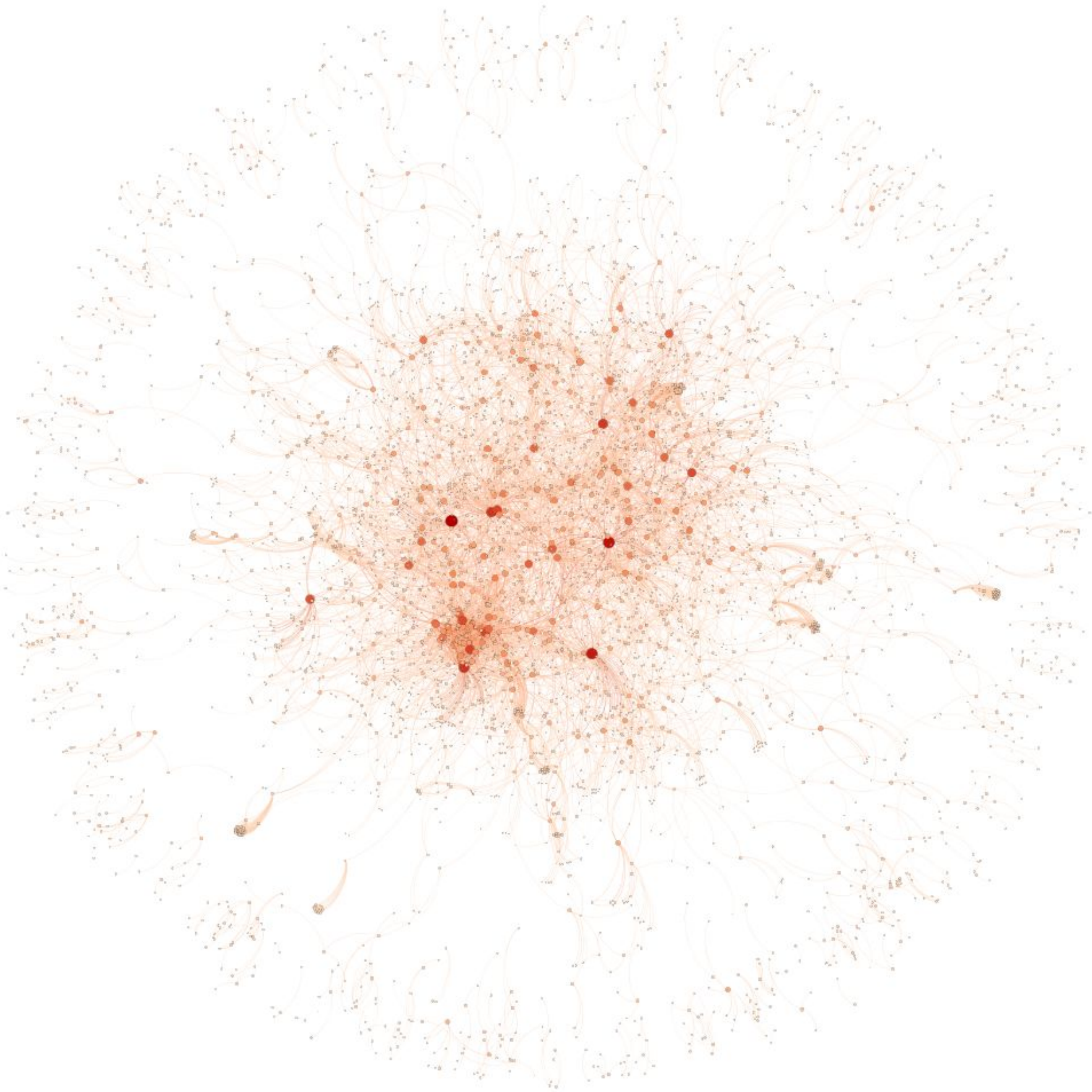
From this plot, we can observe that there is a big portion of nodes with a clustering coefficient near 1 (strongly connected components) and another big number of nodes with coefficient near 0 (hubs).

This observation suggests again our network is scale-free.



-- Number of triangles in which a particular node participate.

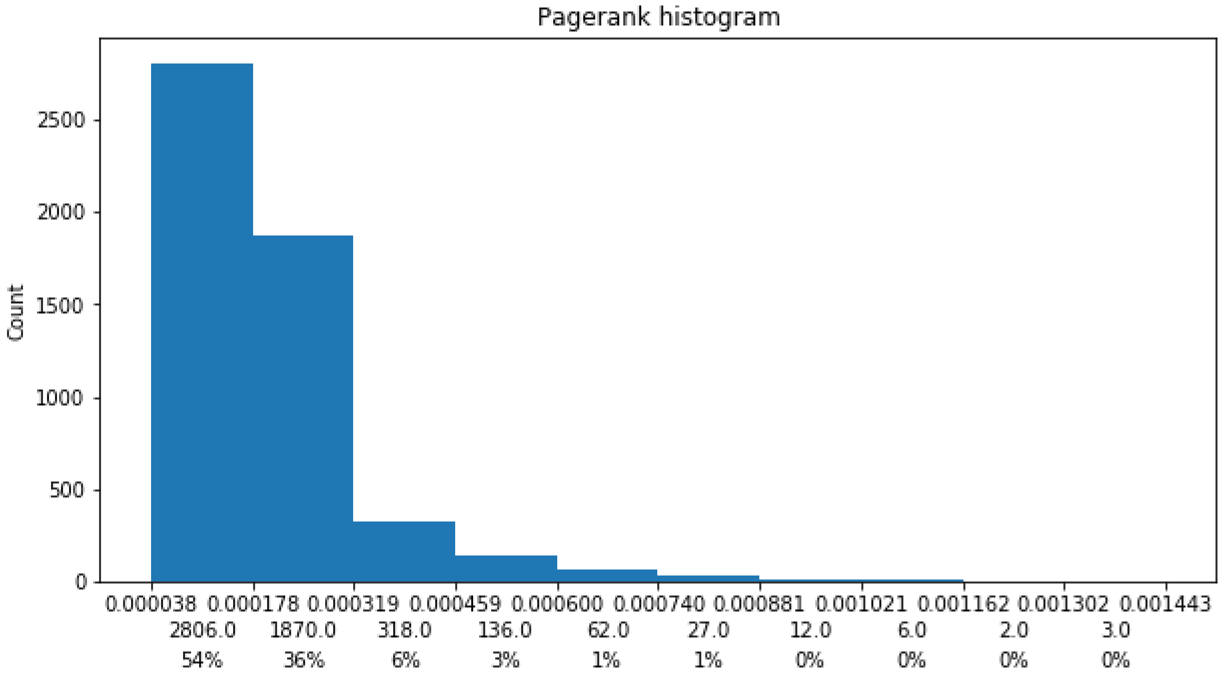
1.5. Pagerank



High rank is assigned to those pages pointed by other important nodes (where the importance refers to the link structure within the network).

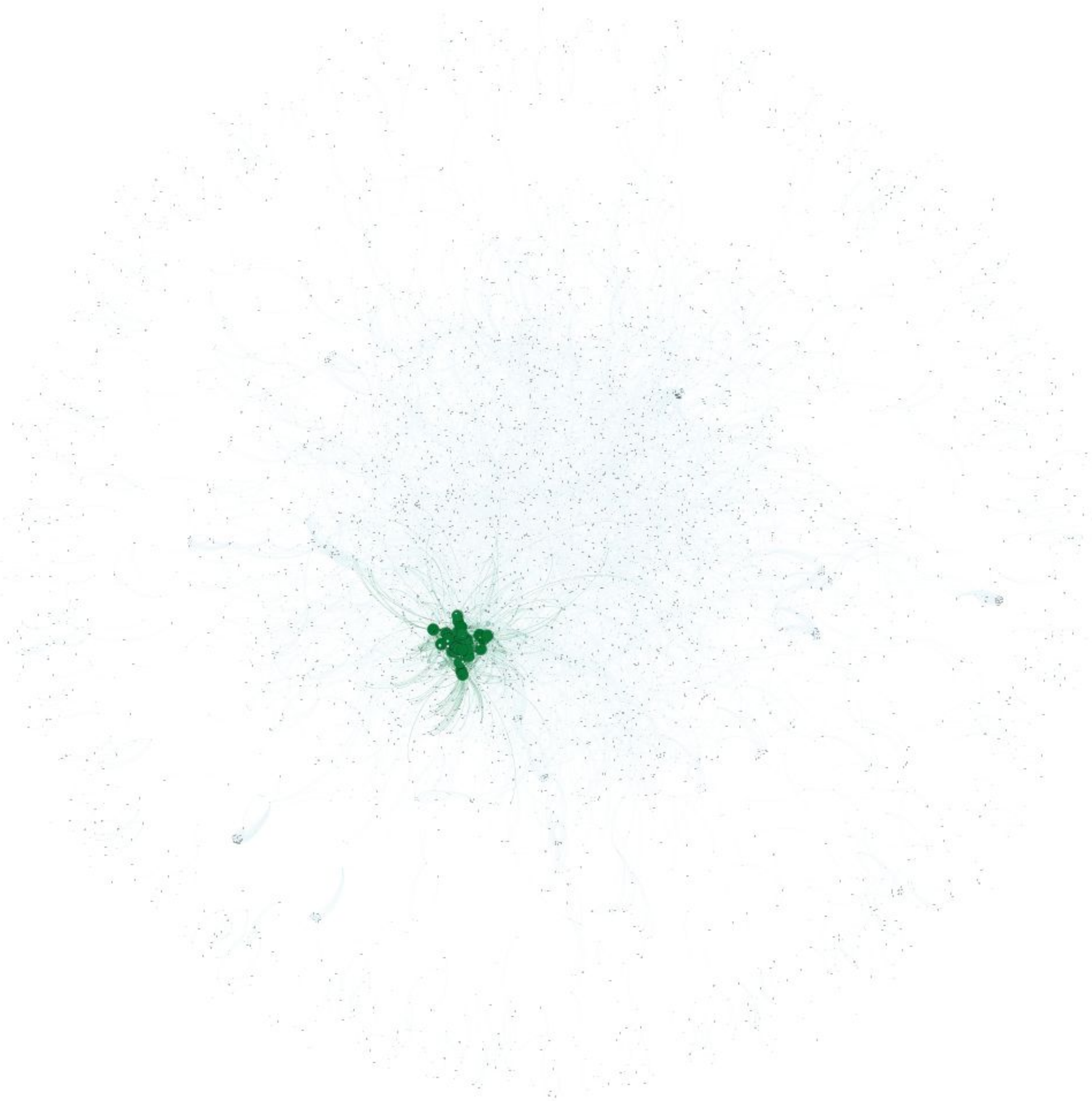
Pagerank distribution:

- average: 0.00
- variance: 1.76e-08
- maximum: ('14265', 0.00)
- minimum: ('4382', 3.8e-05)
- median: ('20116', 0.00)



The function compute the PageRank algorithm with damping factor equal to 0.85 (like Google matrix).

1.6. HITS

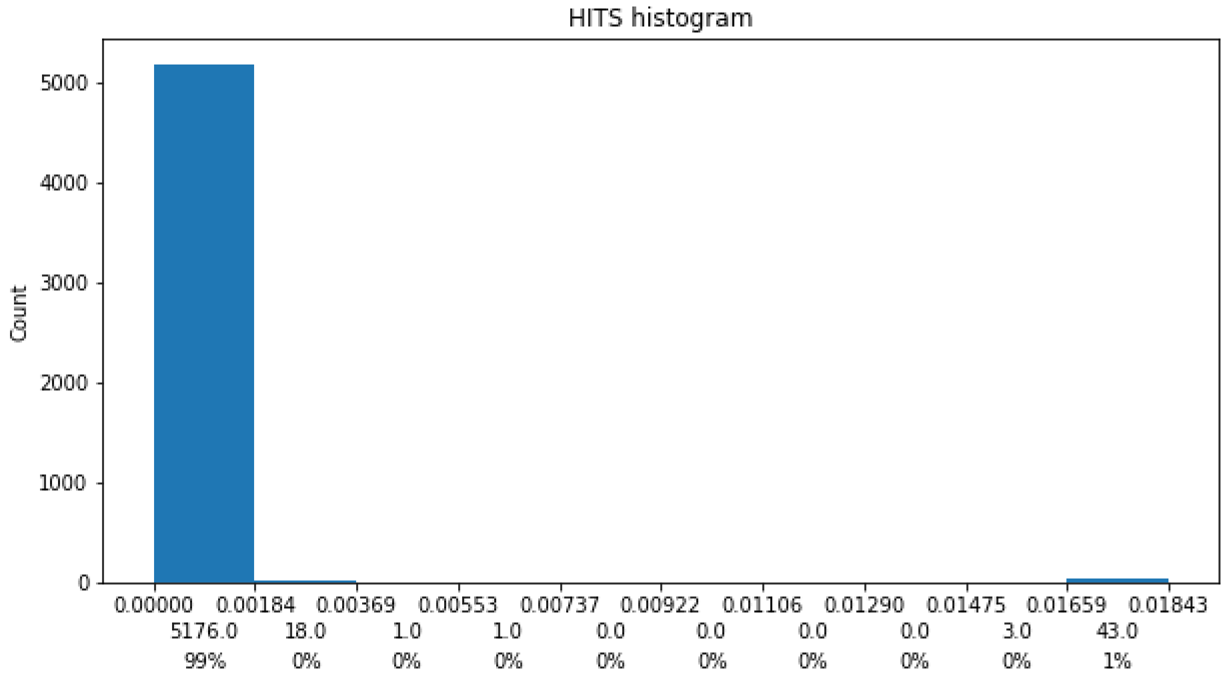


The HITS algorithm computes two numbers for a node. Authorities estimate the node value based on the incoming links, while Hubs estimates the node value based on outgoing links.

However, since the graph we're analysing is undirected, these two values coincide.

HITS distribution:

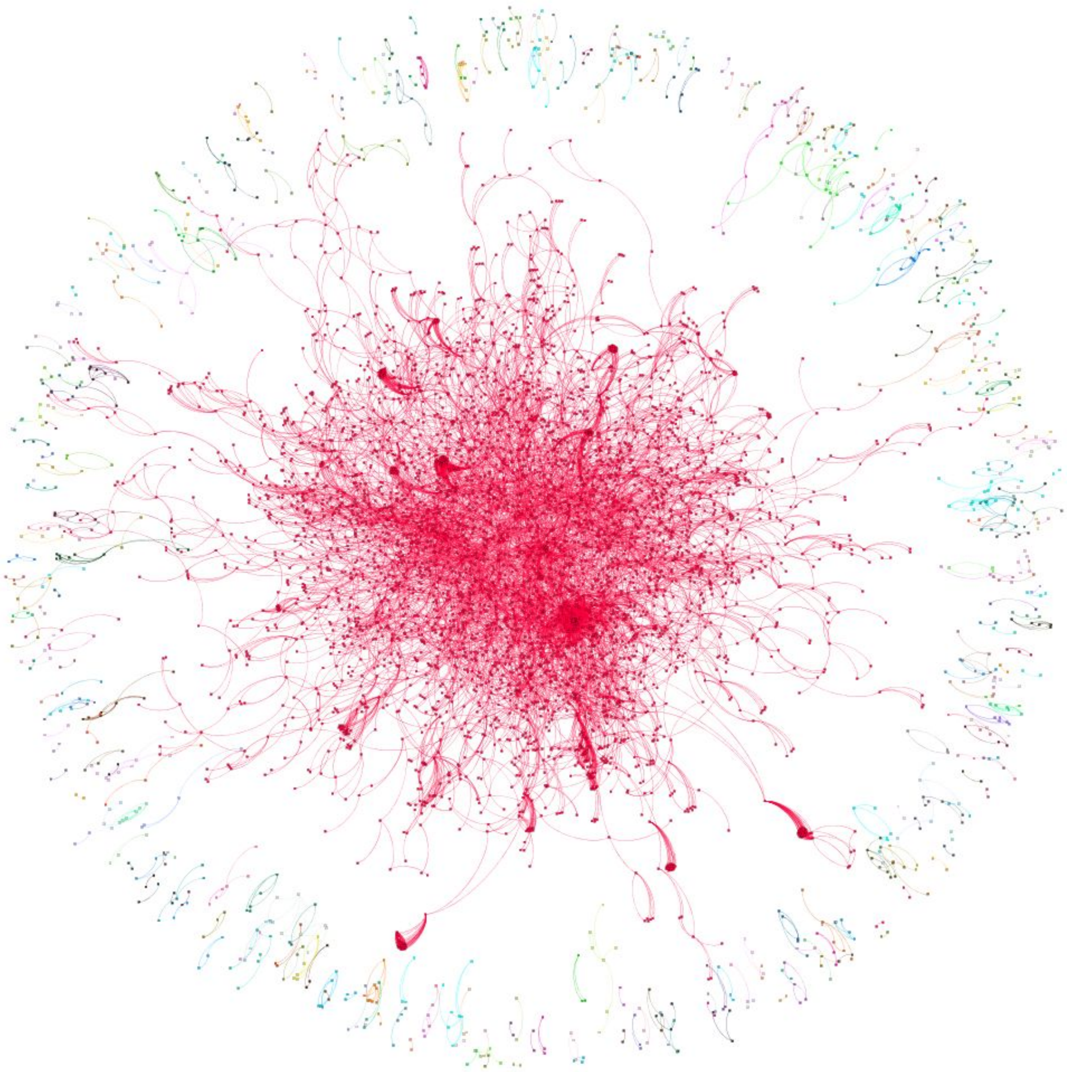
- average: 0.00
- variance: 2.64e-06
- maximum: ('21012', 0.02)
- minimum: ('16470', 0.0)
- median: ('8215', 2.73e-08)



The emergence of hubs is a consequence of a scale-free property of networks: while hubs cannot be observed in a random network, their emergence is expected in scale-free networks. In fact, in the former case, the degree k is comparable for every node (it is therefore not possible for hubs to emerge); on the other hand, in the latter, few nodes have a high degree k while the others yield a small number of links.

2. Graph level measurements

2.1. Giant Component



The Giant component covers 79.32% of the whole graph.

Here the average degree distribution follows the intuition of the random network structure:

$$1 < \langle k \rangle < \ln(n) \mapsto 1 < 5.53 < 8.56,$$

indeed we can observe a giant component $[1 < \langle k \rangle]$and several not connected clusters $[\langle k \rangle < \ln(n)]$.

2.2. Communities detection

2.2.1. Girvan-Newman algorithm

The Girvan-Newman algorithm is an example of a divisive hierarchical clustering; in general, the problem of clustering applied to the graph theory is the possibility to uncover a community structure in a network whose algorithm run time grows polynomially with the graph size.

It is defined as a divisive procedure in the sense that it iteratively removes links connecting nodes that possibly belong to different communities: starting from a single connected cluster we end up eventually breaking the network into isolated aggregates.

The steps are:

1. The betweenness of all existing edges in the network is calculated first.
2. The edge with the highest betweenness is removed.
3. The betweenness of all edges affected by the removal is recalculated.
4. Steps 2 and 3 are repeated until no edges remain.

Obviously, at each iteration we end up with a partition candidate for the optimal communities structure on the network; this iterative break of the graph can be represented by a dendrogram composed by several layers, and at each level, a possible cut will output a valid partition.

Following the algorithm we choose the division that retains the maximum modularity, defined as the fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random:

$$M_c = \frac{1}{2L} \sum_{(i,j)\in C_c} (A_{ij} - p_{ij}) \Rightarrow \cdots \Rightarrow M = \sum_{c=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right],$$

where L_c is the total number of links within the community C_c, L is the total number of links and k_c is the total degree of the nodes in this community.

This definition follows the intuition that randomly wired networks lack an inherent community structure.

2.2.2. Clauset-Newman-Moore greedy modularity maximisation

Another kind of processes relies on the assumption that for a given network the partition with maximum modularity corresponds to the optimal community structure; these are defined as greedy modularity algorithms and the Clauset-Newman-Moore is perhaps the most common.

Simply, the procedure joins pairs of communities if the move increases the partition's modularity; in particular, the algorithm follows these steps:

1. Assign each node to a community of its own, starting with N communities of single nodes.
2. Inspect each community pair connected by at least one link and compute the modularity difference ΔM obtained if we merge them. Identify the community pair for which ΔM is the largest and merge them. Note that modularity is always calculated for the full network.
3. Repeat Step 2 until all nodes merge into a single community, recording M for each step.
4. Select the partition for which M is maximal.

2.3. Communities evaluation

Other than modularity, evaluation of a possible graph partition is carried out keeping track of two other possible common measures:

1. *Coverage*: perhaps the simplest index realising a quality measure of a possible graph partition, keeps track of the fraction of the intra-community edges over all the existent for a cluster c:

$$C(c) = \frac{\#\{(i,j) \in E \mid i,j \in c\}}{\#\{(i,j) \in E \mid i \in c \vee j \in c\}},$$

where E is the edges set for the graph; this measure can be easily extended in order to evaluate the whole partition p:

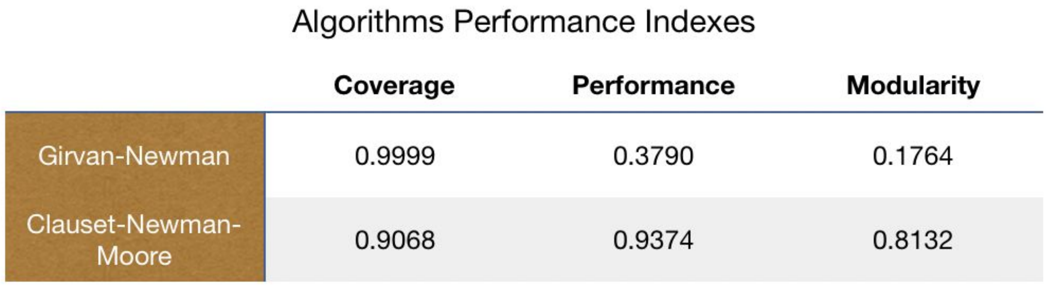
$$Cov_p = \frac{\sum_c^p \#\{(i,j) \in E \mid i,j \in c\}}{\#E}$$

Intuitively, coverage values close to 1 correspond to a good graph partition; however, if we think of a clustering composed by just one community, this will obviously yield the maximum value (so the major drawback of this index is that it will always prefer a partition composed by few groups).

2. *Performance*: It is defined as the fraction of node pairs, that are clustered correctly, i.e. those connected node pairs that are in the same cluster and those non-connected node pairs that are separated by the clustering; for a partition p:

$$P_p = \frac{\#\{(i,j) \in E \mid \varphi(i) = \varphi(j)\} + \#\{(i,j) \notin E \mid \varphi(i) \neq \varphi(j)\}}{n(n-1)/2},$$

where $\varphi(i)$ return the cluster's index of the partition p; the drawback of performance is that in sparse networks, which most real-world networks indeed are, the second addend in the numerator clearly dominates the formula, forcing rather fine clusters.



Algorithms Performance Indexes

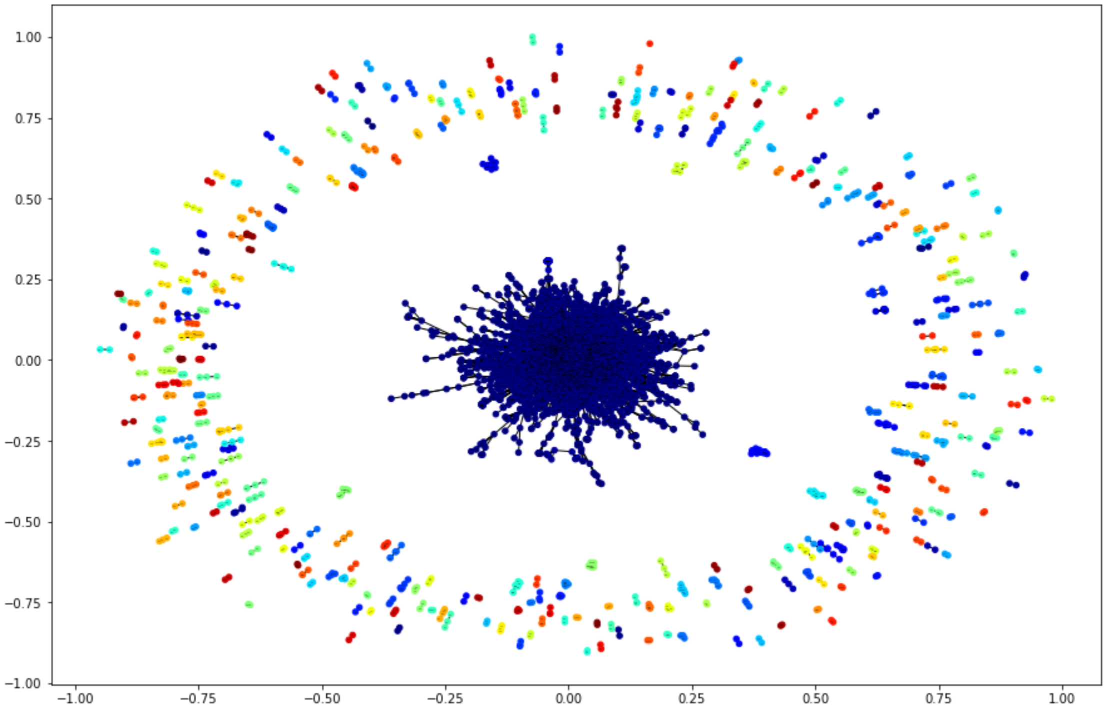
	Coverage	Performance	Modularity
Girvan-Newman	0.9999	0.3790	0.1764
Clauset-Newman-Moore	0.9068	0.9374	0.8132

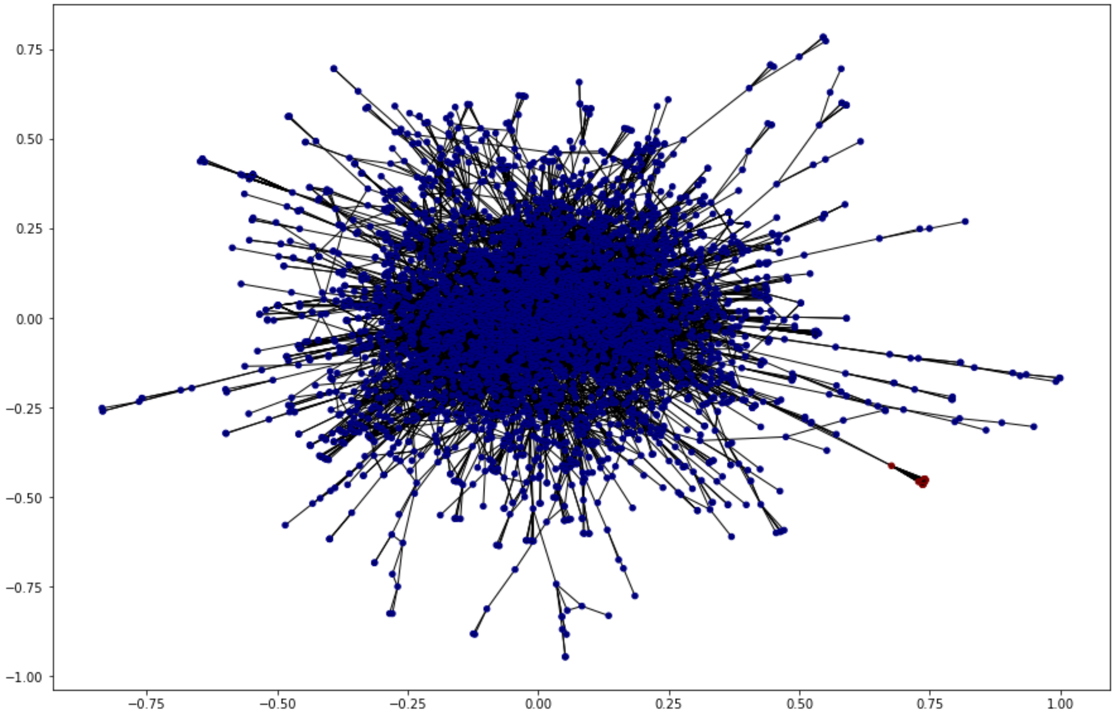
Comparing the two, we can observe that the greedy algorithm try to partition the giant component, thus resulting in a lower coverage index (since we clearly obtain more clusters), on the other hand, it receives a greater modularity thanks to the algorithm definition, as we could have expected a priori.

Since the graph is pretty sparse (14496 out of the 13736661 possible nodes connections), as commented above, finer partitions yield greater performance values, as the C-N-M algorithm does respect to the Girvan-Newman.

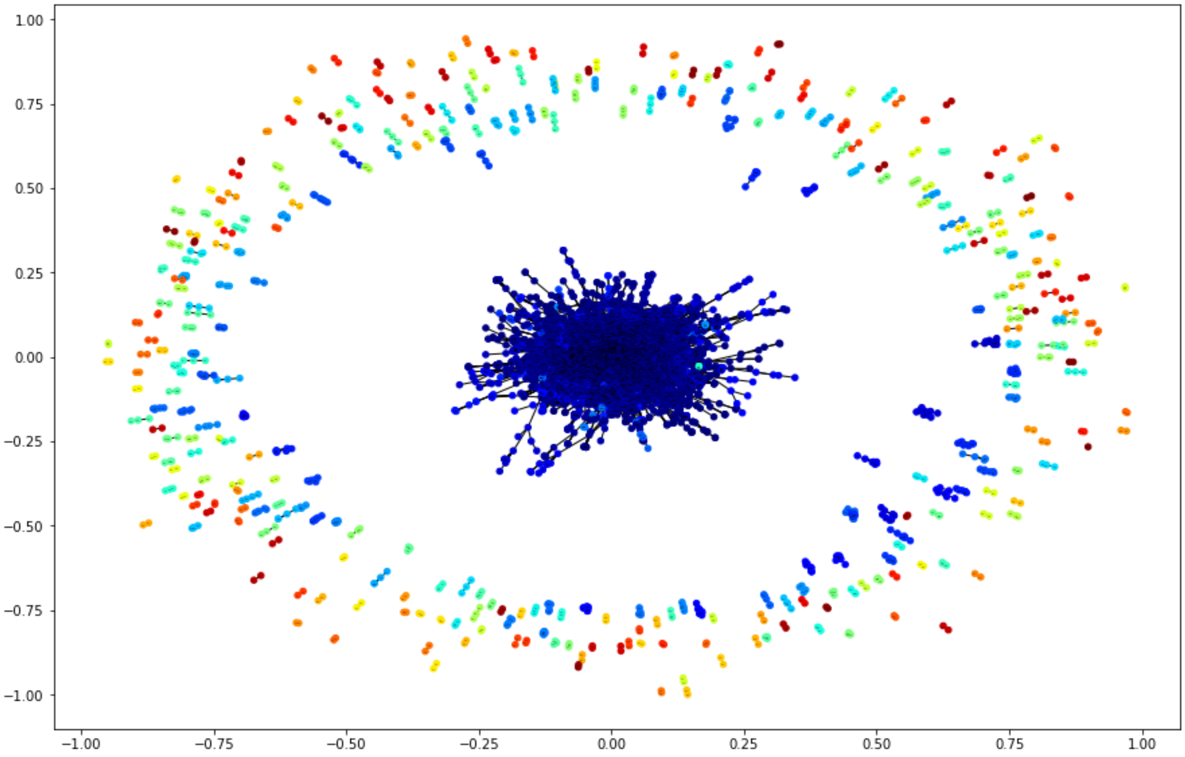
2.4. Communities visualisation

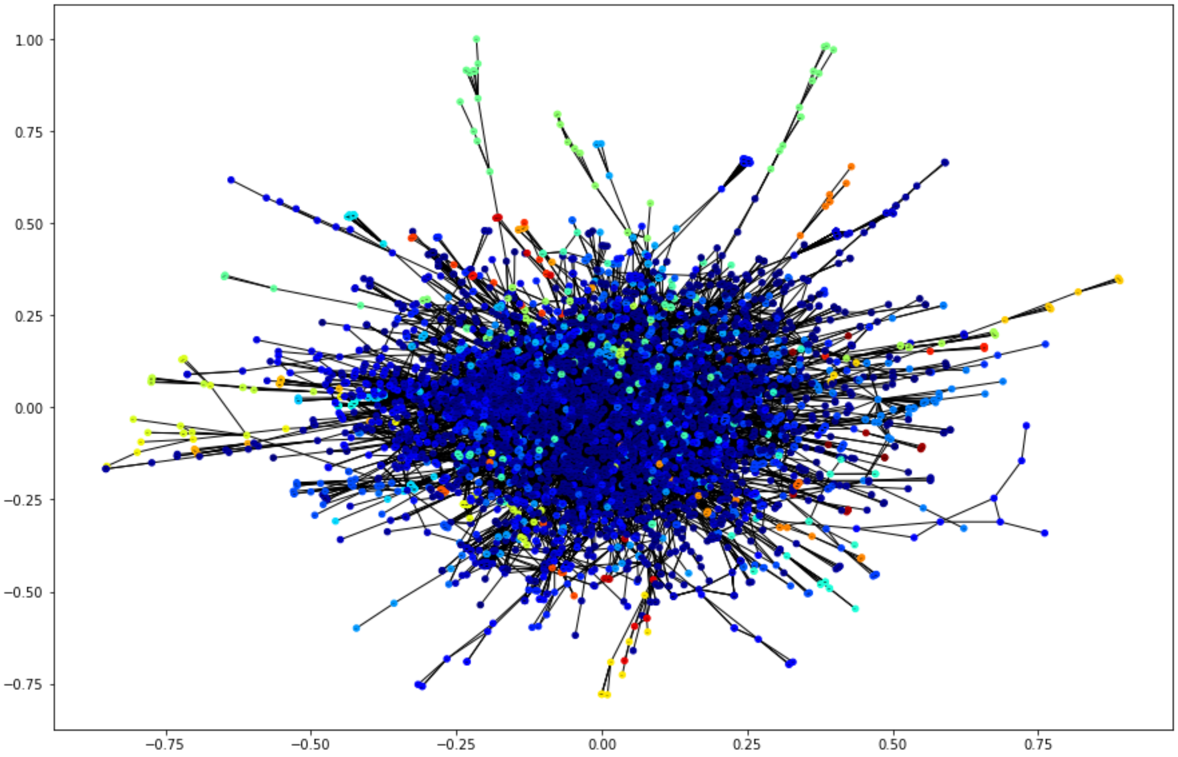
2.4.1. Girvan-Newman





2.4.2. Clauset-Newman-Moore greedy modularity maximisation



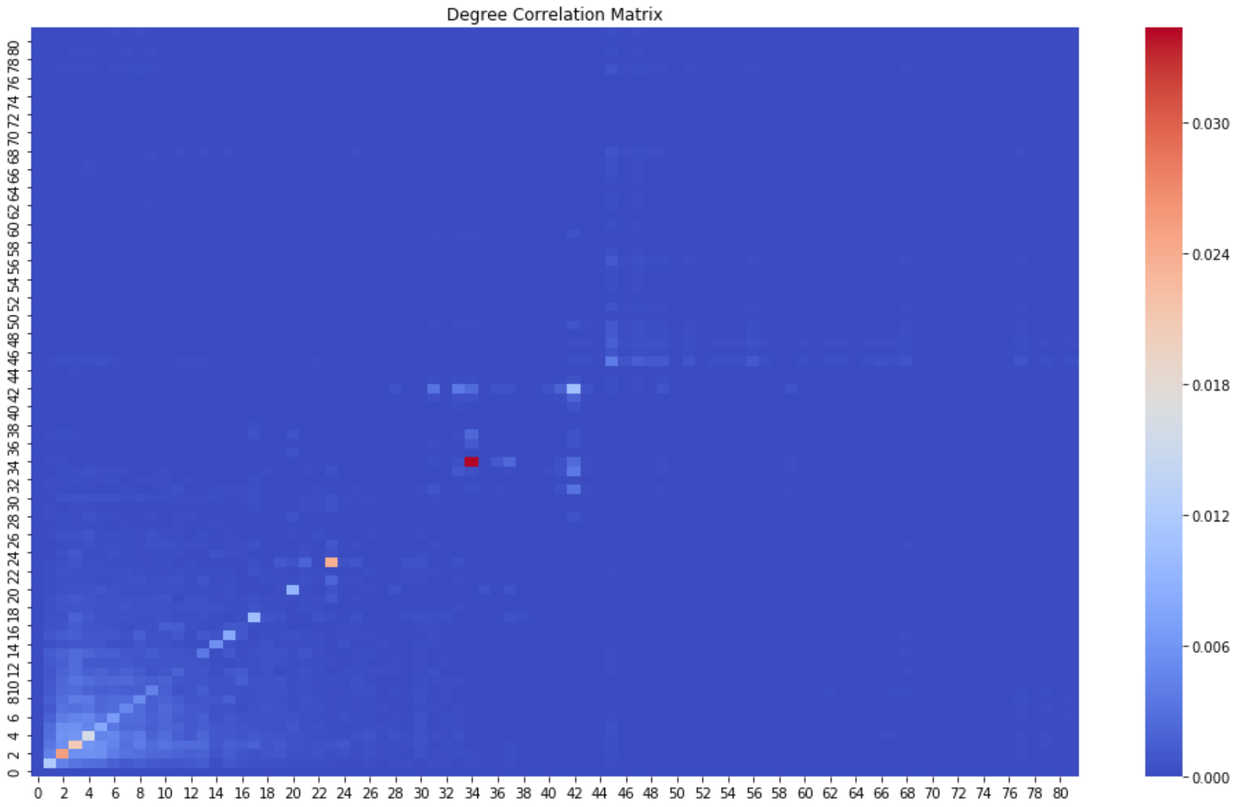


2.5. Measures summary of the whole graph and giant component

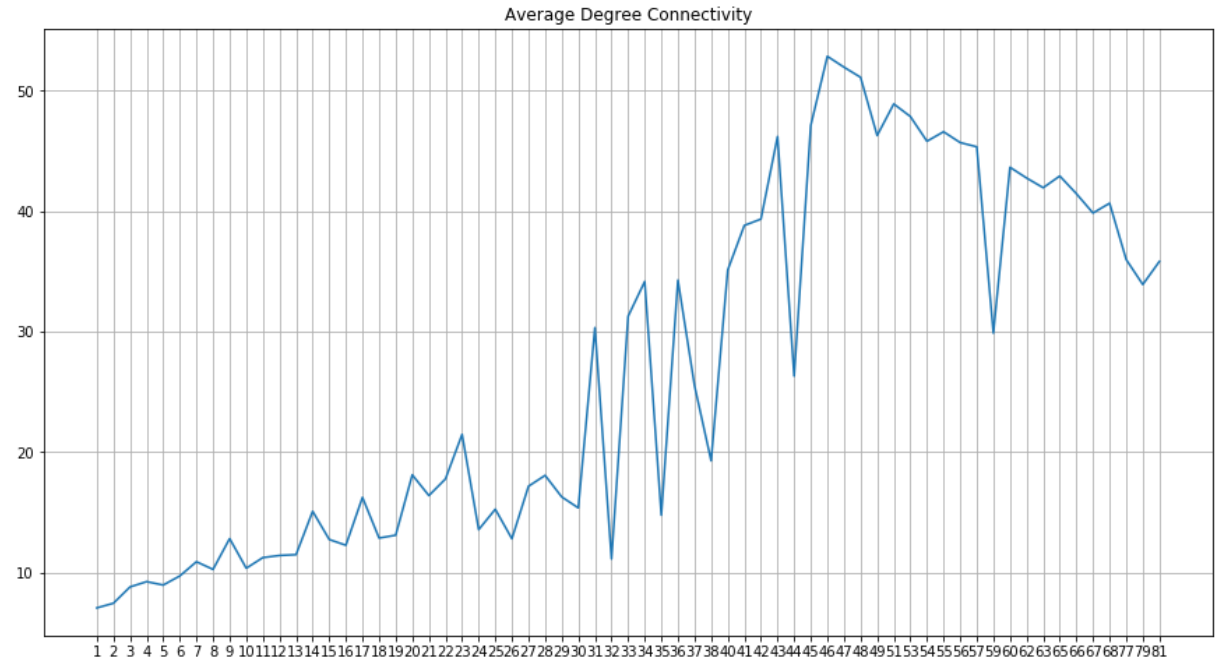
2.5.1. General statistics of the whole graph

- Average degree: 5.5307
- Density: 0.0010
- Diameter: ∞
- Average Path Length: ∞
- Average Clustering Coefficient: 0.5296
- Transitivity: 0.6298
- Assortativity: 0.6592

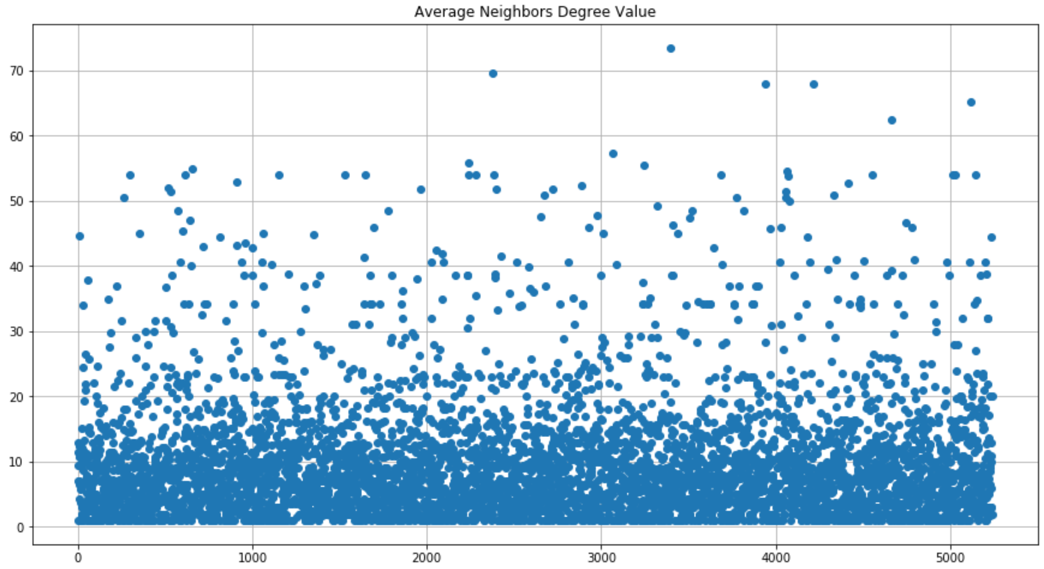
2.5.2. Assortativity measures of the whole graph



The hubs in the heatmap tend to link to each other and avoid linking to small-degree nodes. At the same time, the small-degree nodes tend to connect to other small-degree nodes; networks displaying such trends are assortative.



Again, as confirmation of the network assortative trend, we can observe an increasing average neighbours degree connectivity as k move far from 0 (high degree nodes have, on average, a high degree neighbourhood as well, with a peak around $k \simeq 46$).

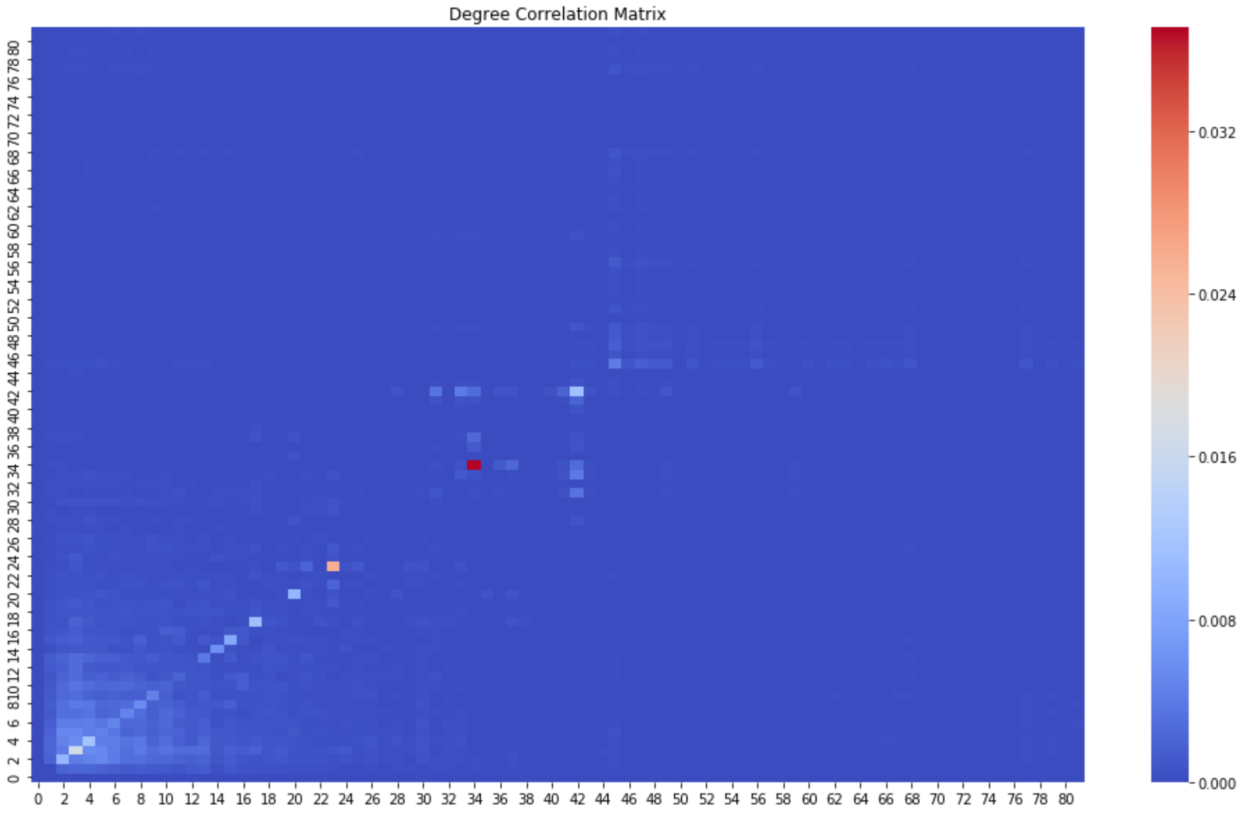


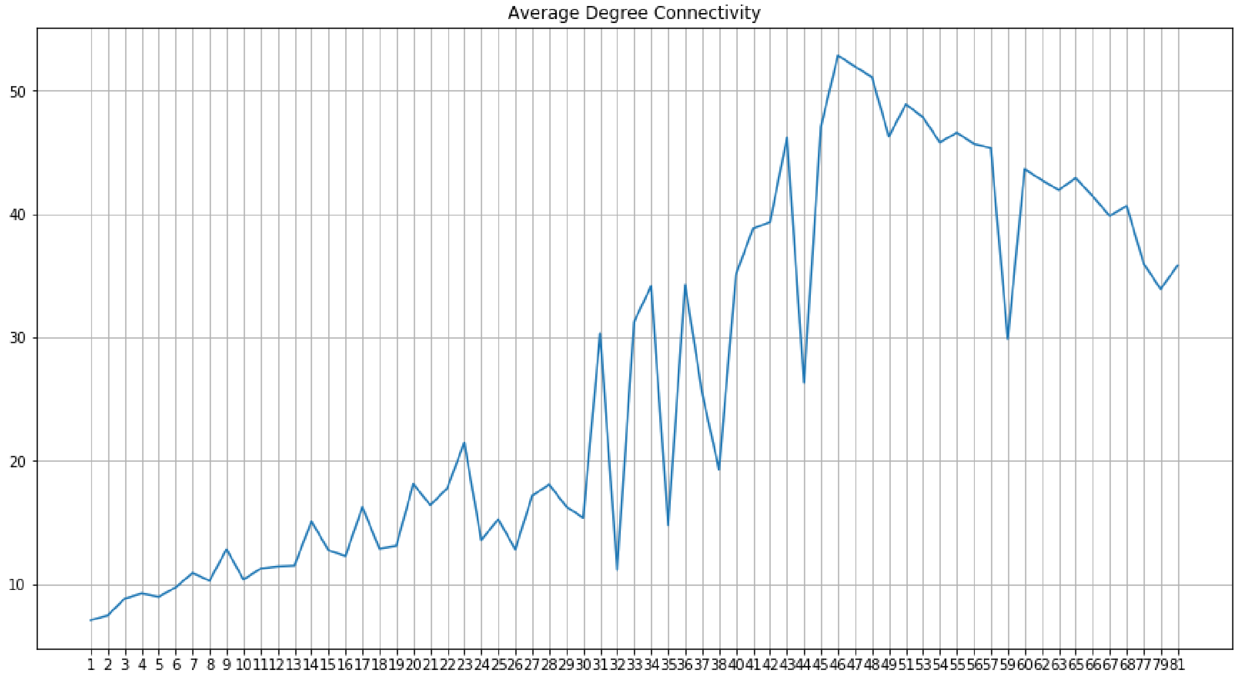
For each node, we plot the average degree of its neighbourhood: the power-law degree distribution is pretty much evident! $\sim 80\%$ of the nodes have an average degree lower than 20, with peaks that surpass 70.

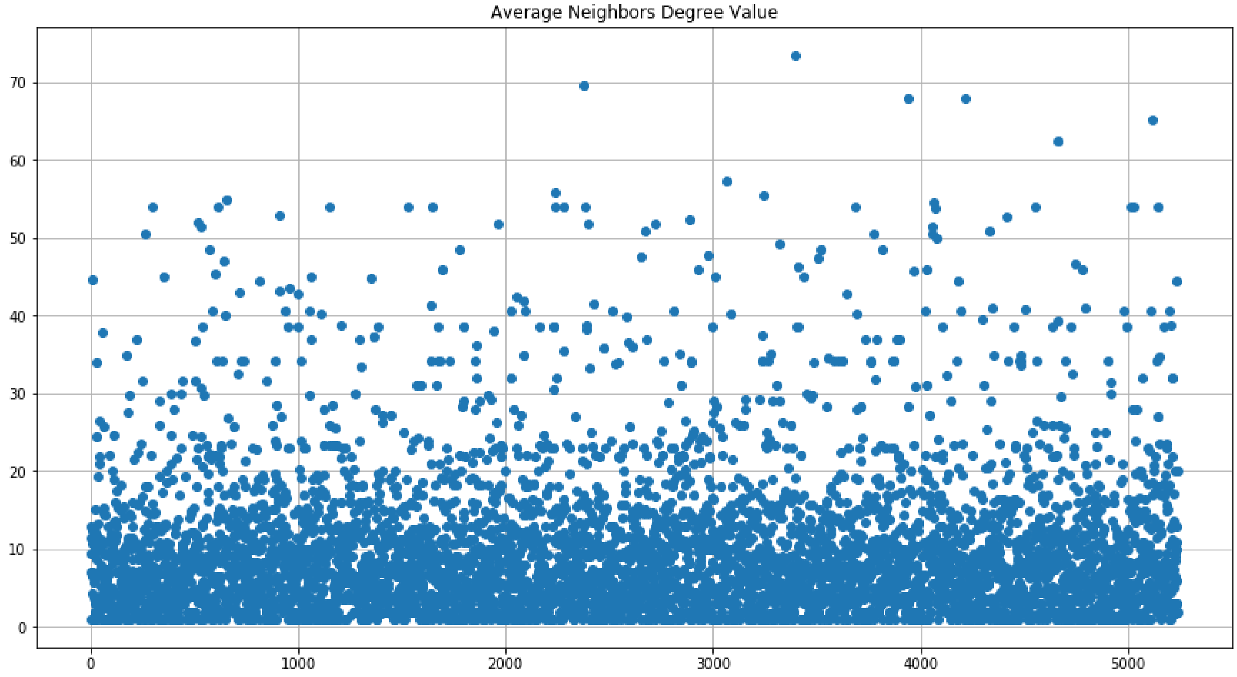
2.5.3. General statistics of the giant component

- Average degree: 6.4589
- Density: 0.0015
- Diameter: 17
- Average Path Length: 6.0494
- Average Clustering Coefficient: 0.5569
- Transitivity: 0.6289
- Assortativity: 0.6390

2.5.4. Assortativity measures of the giant component







Since the network does not show any meaningful components other than the giant one, it easily conceivable that the plots would look like pretty much the same (the periphery consists only of isolated $2 \sim 4$ nodes clusters.

3. Questions

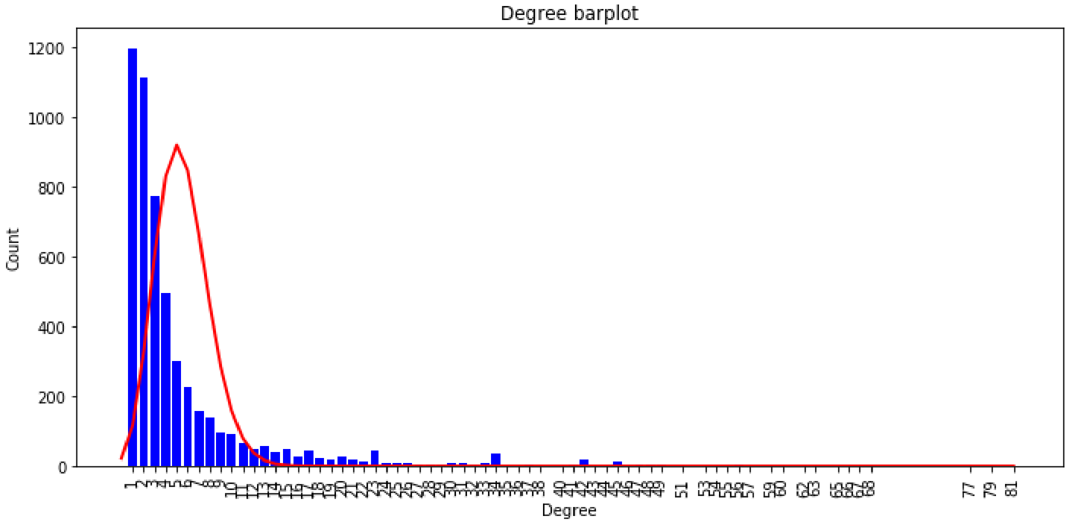
3.1. Does the graph have the same characteristics of a random or a power-law network?

Transitivity: 0.6298

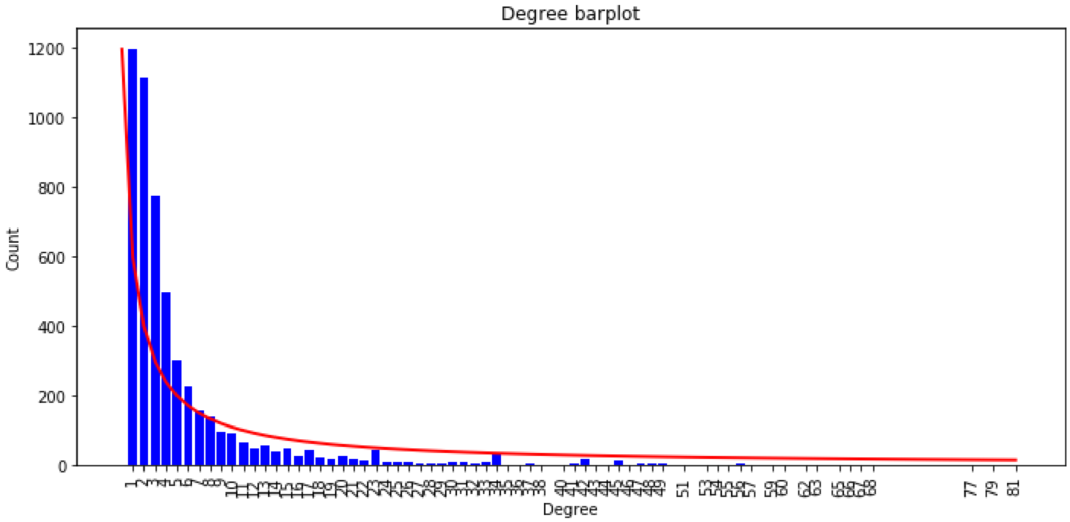
Clustering: 0.5296

Average degree: 5.5307

→ Random Graph Testing



→ Scale-free Testing



3.2. Which are the most important nodes, with respect to a given centrality measure?

- Best 10 nodes in terms of Degree distribution are:

['21012 (81)', '21281 (79)', '22691 (77)', '12365 (77)', '6610 (68)', '9785 (68)', '21508 (67)', '17655 (66)', '2741 (65)', '19423 (63)']

- Best 10 nodes in terms of Betweenness are:

['13801 (0.04)', '9572 (0.03)', '14599 (0.03)', '7689 (0.02)', '13929 (0.02)', '5052 (0.02)', '14485 (0.02)', '2710 (0.02)', '14265 (0.02)', '17655 (0.02)']

- Best 10 nodes in terms of Closeness are:

['13801 (0.19)', '14485 (0.19)', '9572 (0.19)', '17655 (0.19)', '2654 (0.19)', '21012 (0.19)', '12545 (0.19)', '25006 (0.19)', '12365 (0.19)', '22691 (0.18)']

- Best 10 nodes in terms of Clustering are:

['5233 (1.0)', '18720 (1.0)', '14982 (1.0)', '24444 (1.0)', '16766 (1.0)', '16770 (1.0)', '16858 (1.0)', '17389 (1.0)', '495 (1.0)', '20008 (1.0)']

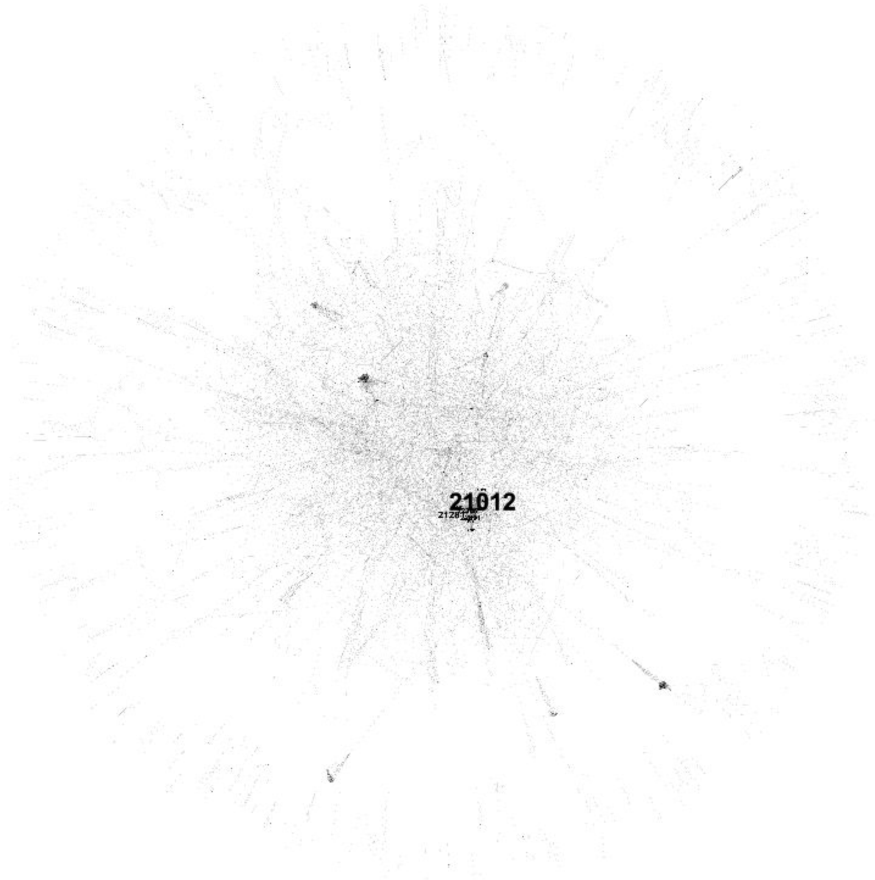
- Best 10 nodes in terms of Pagerank are:

['14265 (0.0)', '13801 (0.0)', '13929 (0.0)', '21281 (0.0)', '9572 (0.0)', '2710 (0.0)', '22691 (0.0)', '21012 (0.0)', '7689 (0.0)', '6264 (0.0)']

- Best 10 nodes in terms of HITS are:

['21012 (0.02)', '2741 (0.02)', '12365 (0.02)', '21508 (0.02)', '9785 (0.02)', '15003 (0.02)', '25346 (0.02)', '7956 (0.02)', '14807 (0.02)', '12781 (0.02)']

The most influential nodes in the network consequently are (node, list appearance):
[('21012', 5), ('22691', 4), ('13801', 4), ('9572', 4), ('21281', 3), ('12365', 3), ('17655', 3), ('7689', 3), ('13929', 3), ('2710', 3), ('14265', 3), ('9785', 2), ('21508', 2), ('2741', 2), ('14485', 2), ('6264', 2)]



3.3. Are the paths short with respect to the size of the network?

If the ratio between the average path length and the diameter is near 0 it means that the short paths are much smaller than the network size; as the opposite, if the value is near 1 it means that the short path is similar to the longest minimum path (geodesic path):

$$PR_G = \frac{\frac{1}{n\cdot(n-1)} \cdot \sum_{i\neq j} d(i,j)}{\max_{i,j} d(i,j)}$$

In the network, the average_distance-diameter ratio is equal to 0.3558.

3.4. Is the network dense?

The total number of possible links in an undirected network is given by $L_{max} = \frac{N(N-1)}{2}$ (N representing the number of nodes), so if we take the ratio between the actual number of edges and the previous quantity, we obtain a decent index that evaluates the sparsity of the network:

$$\frac{L}{L_{max}} = \frac{2L}{N(N-1)}$$

The graph, actually, is mostly sparse: it suggests again that the network is scale-free (within the same network few extremely linked nodes coexist with many sparsely linked ones); in fact, the following hold:

$$L \ll L_{max} \rightarrow$$

$$L = 14\,496 \ll L_{max} = \frac{5\,242(5\,242-1)}{2} = 13\,736\,661$$

A graph is defined as dense when the number of links approximately reaches the number of nodes squared:

$$L \approx L_{max} = \frac{N(N-1)}{2} \sim N^2$$

In the example:

$$L = 14\,496 \not\approx 27\,478\,564 = N^2$$