

Social Network Analysis Project

Luigi Maiese – 0622900080 – l.maiese1@studenti.unisa.it

Mario Cerino – 0622701191 – m.cerino8@studenti.unisa.it

Alberto Vitiello – 0622701186 – a.vitiello62@studenti.unisa.it

Arcangelo Palma – 0622701208 – a.palma27@studenti.unisa.it

Contents

[Midterm Project 3](#_Toc77585438)

[Exercise 1 3](#_Toc77585439)

[Task 3](#_Toc77585440)

[Solution 3](#_Toc77585441)

[Exercise 2 4](#_Toc77585442)

[Task 4](#_Toc77585443)

[Solution 4](#_Toc77585444)

[Performance Analysis 4](#_Toc77585445)

[Exercise 3 6](#_Toc77585446)

[Task 6](#_Toc77585447)

[Solution 6](#_Toc77585448)

[Exercise 4 9](#_Toc77585449)

[Task 9](#_Toc77585450)

[Solution 9](#_Toc77585451)

[Final Project 10](#_Toc77585452)

[Exercise 1 10](#_Toc77585453)

[Task 10](#_Toc77585454)

[Solution 10](#_Toc77585455)

[Exercise 2 12](#_Toc77585456)

[Task 12](#_Toc77585457)

[Solution 12](#_Toc77585458)

[Exercise 3 26](#_Toc77585459)

[Task 26](#_Toc77585460)

[Solution 27](#_Toc77585461)

# Midterm Project

## Exercise 1

### Task

Consider the “Facebook Large Page-Page Network” dataset available at the Stanford Large Network Dataset Collection (SNAP) (https://snap.stanford.edu/data/facebook-large-page-page-network.html). Note that network edges are provided in the file musae\_facebook\_edges.csv contained in the zip file linked in this page. Analysing this file, you can see that nodes of the network can be partitioned in four categories: politicians, governmental organizations, television, shows and companies. We will refer to this partition as the real clustering. You are required to cluster nodes of the network in at least 4 clusters using each of the partition algorithms seen in class: hierarchical, k-means, Girman-Newmann (betweenness-based clustering), spectral. Note that the network is very large and the naive implementations of these algorithms may be very expensive. Hence, you are required to optimize these algorithms (by sampling, parallelism, and adhoc optimizations) to make their running times feasible. Compare the clustering obtained through each of your algorithms with respect to the real clustering given in the file musae\_facebook\_target.csv. Discuss the trade-off between precision and running time of each of your proposed implementations.

### Solution

For exercise 1 the network has been clustered using the following methods:

1. Hierarchical:
   1. Naïve
   2. Custom ad-hoc optimization
2. K-Means
   1. Naïve
   2. Custom ad-hoc optimization
3. Betweenness:
   1. Naïve
   2. Parallel
4. Spectral:
   1. Naïve
   2. Parallel

## Exercise 2

### Task

Return the top 500 nodes (that are approximatively the top 2%) of the Facebook Large Page-Page Network according to each of the following centrality measures: degree, closeness, betweenness, PageRank, HITS. For the first three measures you can use the algorithms presented in class. For PageRank and HITS algorithms you have to provide both a naive and a parallel implementation. As in the previous exercise, you have to optimize your algorithms to make their running times feasible on a very large network. For each of the required centrality measures describe which is its best implementation, by taking into account both the running time and the precision. In particular, for the measures involving the choice of some parameters, such as PageRank and HITS, discuss of the best choice of parameters. Compare the results of the different algorithms and discuss about the similarities and the differences among the returned outcomes.

### Solution

For exercise 2 the network has been analysed using the following centrality measures:

1. Degree:
   1. Naïve
2. Closeness
   1. Naïve
   2. Parallel
3. Betweenness:
   1. Naïve
   2. Parallel
4. PageRank:
   1. Naïve
   2. Vectorized
   3. Networkx
5. HITS:
   1. Naïve
   2. Parallel

### Performance Analysis

The heuristic used for the evaluation of each centrality measure follows these rules:

1. The ground truth of each centrality measure is represented by the result of the naïve version.
2. Each optimization is evaluated with respect to:
   1. Running time.
   2. Precision compared to the ground truth.

For this purpose, we chose a rule that enabled us to make this decision, in particular we analysed the results following two parameters:

* **Similarity Rate**: this parameter represents the percentage of the nodes in the top 500 of the ground truth that are also in the top 500 of the optimized version
* **Equality Rate**: this parameter represents the percentage of the nodes in the top 500 of the ground truth that are in the same position respect to the top 500 of the optimized version

## Exercise 3

Discussion about exercise 3

### Task

Consider the following scenario. A restaurant is evaluated by a reviewer with respect to three features: Food, Service, and Value. For each of these features the reviewer can assign from 0 to 5 points. Observe that not all the restaurants can be evaluated with respect to all the features. Indeed, even if all restaurants must be always evaluated on food. a restaurant that only offers take-away service cannot be evaluated about service. Similarly, a restaurant that only runs as company canteen cannot be evaluated about value. The Michelin guide must assign a score to each restaurant. Three scores are possible: one star, two stars, three stars. This year, the Michelin guide’s principal has decided to not use his expensive team of experts to evaluate the restaurants, but to run an algorithm that, taken in input the scoring assigned by the experts in the past, and some new reviews, assigns the scores. However, the algorithm must avoid that a restaurant “with service” will receive a higher score than when it declares to be “only take-away”. Similarly, it must avoid that a restaurant “a la carte” will receives a higher score when it declares to be a “company canteen”. Provide a classifier for the Michelin guide’s principal that satisfies all the required features. You must convince the principal that your approach satisfies the required features, either by providing a formal proof, or by running massive experiments showing that the required features are (almost) always metre.

### Solution

Our solution for exercise three is the implementation of the MinCut algorithm that can make a classification of the stars assigned to a restaurant with respect to its features.

For this purpose, we first implemented three algorithms for the creation of three different kinds of dataset. As the task exploits, each restaurant is evaluated with respect to food and can be also evaluated with respect to service and value (if the restaurant provides those features), so the dataset is created assigning for each restaurant a vote from 0 to 5 if a feature is provided, -1 if not.

* **Dataset 1**: the first dataset is created assigning for each restaurant a number of stars considering the max value among its food, service and value with respect to a random probability.

def max\_based\_dataset(number\_of\_iterations)

* **Dataset 2**: the second dataset is created assigning for each restaurant a number of stars considering the average value among its food, service and value with respect to a random probability.

def average\_based\_dataset(number\_of\_iterations)

* **Dataset 3**: the third dataset is purely random.

def totally\_random\_dataset(number\_of\_iterations)

MinCut algorithm takes in input a dictionary that contains the a priori probability of each tuple [food, service, value] to receive a given star number based on statistical average.

def probability\_computation(dataset)

MinCut algorithm is a binary classifier; in this scenario we have three possible classifications (1 star, 2 stars, 3 stars) so to obtain a MinCut that can do a three-class classification we followed these steps:

def mincut\_algorithm(probability\_dict):

* **Step 1**: run the naïve MinCut algorithm on the whole graph, so giving to the algorithm the complete graph also containing the dictionary of probabilities calculated before. This MinCut returns a **partition** that contains two groups:
  + each tuple [food, service, value] that have a star score equal to one.
  + each tuple [food, service, value] that have a star score equal to two or three.
* **Step 2**: run the naïve MinCut algorithm on a partial graph composed by every tuple [food, service, value] contained into the second partition obtained from **Step 1**. This MinCut returns a **partition** that contains two groups:
  + each tuple [food, service, value] that have a star score equal to two.
  + each tuple [food, service, value] that have a star score equal to three.
* **Step 3**: creation of a dictionary that contains the results of MinCut classification.

We need to highlight what happens in the first two steps:

* **Step 1:**
  + each tuple [food, service, value] is linked to “s” with the probability to receive only one star.
  + for each tuple [food, service, value] with at least one feature between service and value, the algorithm creates the edges related to hidden features. This is crucial to guarantee the Truthfulness.
  + each tuple [food, service, value] is linked to “t” with the probability to receive more than one star.
* **Step 2**:
  + for each tuple [food, service, value] the MinCut takes the biggest probability between the chance of getting two or three stars.
  + each tuple [food, service, value] is linked to “s” with the probability to receive two stars.
  + for each tuple [food, service, value] with at least one feature between service and value, the algorithm creates the edges related to hidden features. This is crucial to guarantee the Truthfulness.
  + each tuple [food, service, value] is linked to “t” with the probability to receive three stars.

The task asked to prove that our solution satisfies the required features using a formal proof or massive experiments. To convince the principal, we implemented a function that checks the rules.

def isTruthful(result):

Also, in the main function of *main\_es\_3.py* there is a benchmark that runs the algorithm several time using all the datasets described above (every dataset is generated several time).

All those experiments return positive results and the average execution time of the MinCut algorithm is **0.018168052832285553s**.

## Exercise 4

Discussion about exercise 4

### Task

Consider the same setting as in the previous exercise but suppose that the principal can accept to trade off precision and efficiency of the classifier with its robustness. Thus, it is accepted that misreporting occurs, but we would like that the cases in which a restaurant has an incentive to misreport its features are as few as possible, but he requires that the classifier be as precise and fat as possible. Provide a classifier to the principal that satisfies these requirements. Motivate your choice by comparing your choice with different alternatives and showing how your choice experimentally outperforms the other alternatives in terms of incentive-compatibility, precision, or performance.

### Solution

Our solution for exercise four is the implementation of two/three models trained for the classification of the restaurants:

* Logistic Regressor
* Linear Regressor

To train, validate and test our models we implemented four algorithms for the creation of four different kinds of dataset.

* **Dataset 1**: the first dataset is created assigning for each restaurant a number of stars considering the weighted average of food, service and value with respect to the assignment of three random coefficients. To avoid the discrimination for missing features, if a restaurant is lack of a feature the algorithm assigns a random value.

def coefficient\_based\_dataset(number\_of\_iterations)

* **Dataset 2**: the second dataset is created assigning for each restaurant a number of stars considering the max value among its food, service and value with respect to a random probability.

def max\_based\_dataset(number\_of\_iterations)

* **Dataset 3**: the third dataset is created assigning for each restaurant a number of stars considering the average value among its food, service and value with respect to a random probability.

def average\_based\_dataset(number\_of\_iterations)

* **Dataset 4**: the fourth dataset is purely random.

def totally\_random\_dataset(number\_of\_iterations)

# Final Project

Introduction to midterm project

## Exercise 1

Discussion about exercise 1

### Task

Implement the following game-theoretic centrality measures:

1. shapley\_degree:

this is the Shapley value for the characteristic function value(C) = |C| + |N(C)|, where N(C) is the set of nodes outside C with at least one neighbour in C;

1. shapley\_threshold(k):

this is the Shapley value for the characteristic function value(C) = |C| + |N(C, k)|, where N(C, k) is the set of nodes outside C with at least k neighbours in C;

1. shapley\_closeness:

this is the Shapley value for the characteristic function value(C) = Σu 1/dist(u, C), where dist(u, C) is the minimum distance between u and a node of C.

Recall that the naive implementation of Shapley value requires a running time that is exponential in the number of nodes of the network. You are instead required to provide a polynomial time algorithm for the above measures. On the e-learning platform you fill find material that will help you in designing and implementing these algorithms.

Implement also the Friedkin-Johnsen (FJ) dynamics, that works as follows:

* each node u has a private belief bu in [0, 1] and a stubbornness value su in [0,1];
* at each time step t each node publicizes an opinion xu(t) in [0,1] where:
  + xu(0) = bu, i.e., the initial opinion is exactly its belief;
  + xu(t) = su bu + (1-su) sumv in N(u) 1/N(u) xv(t-1), i.e., the opinion at time t is a weighted average of the private belief and of the opinion publicized by its neighbours at the previous step.

Does these dynamics converge to a stable state (i.e., a state in which no agent updates her opinion – you may assume a finite precision for opinion of at most 5 decimal digits)? Provide either a formal proof or experimental evidence for your answer.

### Solution

Proposed solution.

## Exercise 2

Discussion about exercise 2

### Task

Consider the network N represented in the file net\_x, that has been generated with one of the network models seen during the course.

You have to analyse the network N and guess which model has been used for creating it. Your guess has to be supported by an appropriate set of experiments to confirm that networks generated with the proposed model have characteristics similar to N (note that you have to guess also the parameters of the model).

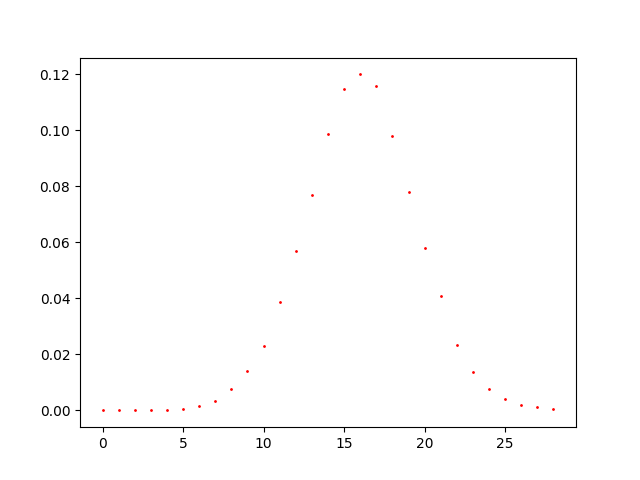
During the discussion of the project, you will be asked to motivate your guess. Motivations may be related to both theoretical properties of the models seen during the course (e.g., “I analysed the provided network and I observed that its node degree distribution follows a power law. Hence, I conclude that it is not possible that the graph has been generated with a model random(n,p).”), and to experimental evidence (e.g., “I generated a lot of random graphs with p = 1/3, and none of them had similar properties as the provided network. Hence I conclude that it is improbable that the graph is random(n, 1/3)”).

A bonus point will be assigned to all the components of the groups that correctly guessed the model (and parameters) used to generate N.

### Solution

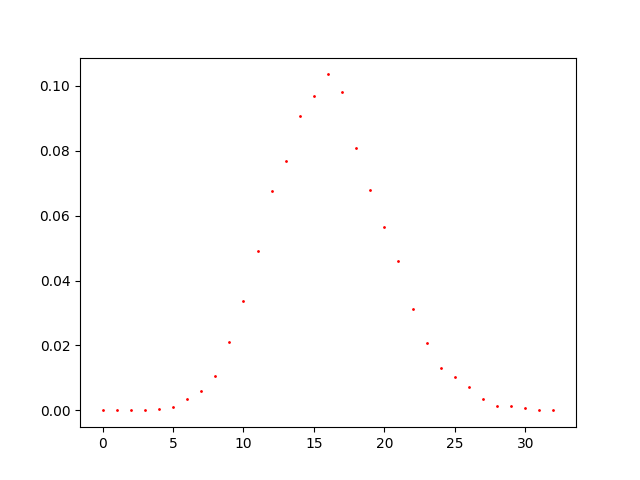
In order to guess the model that has generated the network 6, we start by plotting a scatterplot of the degree distribution, where on the x-axis we put the number of degree and on the y-axis the number of nodes having that degree.

The degree distribution has the following shape:



This figure suggests to us a piece of information. The shape is Gaussian-like, it has mean 16, but it is slightly skewed on the left, with a shift on the right part. Just looking the figure, we can guess that the process that has generated the graph is not a power low model, since the plot doesn’t show the classical long tail.

Since the shape is Gaussian-like, we start our analysis hypothesizing that a random process generated the graph, since the random graphs have a Poisson distribution, that is near to a Gaussian when gamma is grater than 1000. According to our thinking, we start a quantitative analysis about the graph. The original graph has 10000 nodes and 80183 edges, but the fact that grips our attention is the amounts of degree. The node showing the maximum degree has 29 as degree, and the median value of the degrees is 16. If a random model generated the graph, the p should be a small value. In order to guess p, we can observe the mean of the Gaussian-like shape. The mean value is 16, and in order to obtain a shape with these properties, with 10000 nodes, we should use a probability p=0.0016. The same number could be the result of the formula proposed by Newmann, where the mean degree of a random graph is c = (n-1)p. After our reasoning, we generated a random graph with p=0.0016 and the result is a shape like this:



The graph generated and used to plot the above figure, has 10000 nodes, 80224 edges and the maximum degree is 35. These values encourage us to guess that a random model generated the graph. The only problem is that a random graph usually has the tails of equal lengths.

So we try to test our thesis by using some quantitative results. Usually a random graph has a low clustering coefficient. It has also a giant component and few outliers and the diameter, in expectation, is log(N), indeed as Newman says, in a random graph with c > 1 there exists a giant component that fills an extensive fraction of the network

A very simple quantity to calculate for the random graph is the clustering coefficient. Recall that the clustering coefficient C is a measure of the transitivity in a network and is defined as the probability that two network neighbors of a vertex are also neighbors of each other. In a random graph the probability that any two vertices are neighbors is exactly the same— all such probabilities are equal to p = c/(n– 1). Hence

C = c/(n-1)

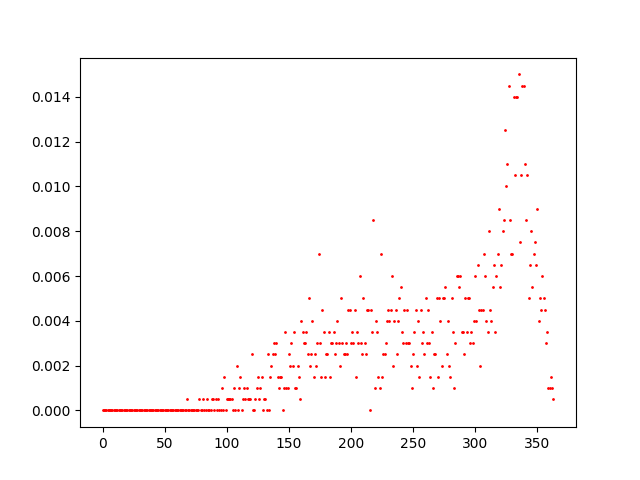
By analysing our networks, the cluster coefficient is 0.07 which is a little number, but isn’t as little as the expectation of a random graph with p=0.0016 and 10000 nodes, which cluster coefficient should be in expectation C=0.0016. This number is confirmed by the mean clustering coefficient of 100 generated networks, which is 0.001545. The fact that the clustering coefficient isn’t as small as we expected, that the original curve is shifted to right with a slightly but considerable skew on the left, suggests us to try other models. Another motivation to declare that this is not a random graph is its diameter. The expectation in diameter of a random graph is log(n), in our case 4. Although the computed diameter for our graph is 6, which is near to 4, the number 6 makes us feel confident the network is a small world, as the expected value from a small world, according to Milgram’s experiment, is just 6.

Discarding power low distribution and random graph, we will try Generalized Watts-Strogatz model

In order to try as much as possible configurations of the params r,k and q, we start by generating graphs of 4000 nodes.

We don’t start with random values, we try to do some reasonings about which parameters could be the most likelihood. Starting with high values for r and/or k could be useless, because each node of net 6 has at maximum 29 degrees. This means that taking a great value for r will increase much the size of the maximum degree. As a proof of our reasoning we report a plot with r=10.

r=10, k=1, q=1

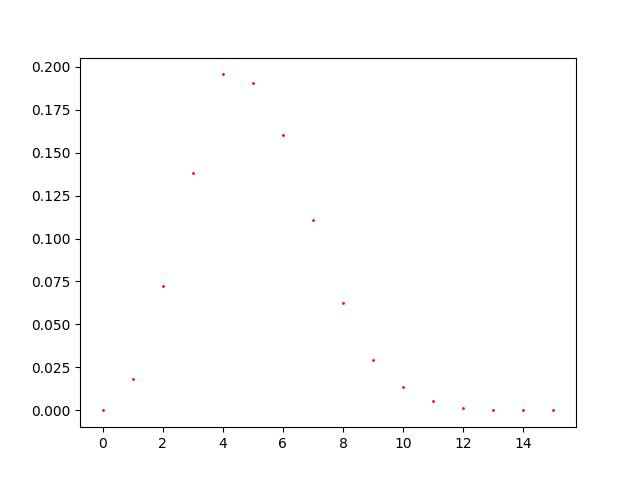


As we can see each node could be a maximum degree of more than 300.

As we can thought, starting by choosing random parameters is not the right way. Although it could be very expensive to guess parameters by random, we can select a range of values for parameters that could be reasonable. For example, since the network has such maximum degree which is very low, we start by setting parameters with r=1, k=1, q=1.

Nodes=4000

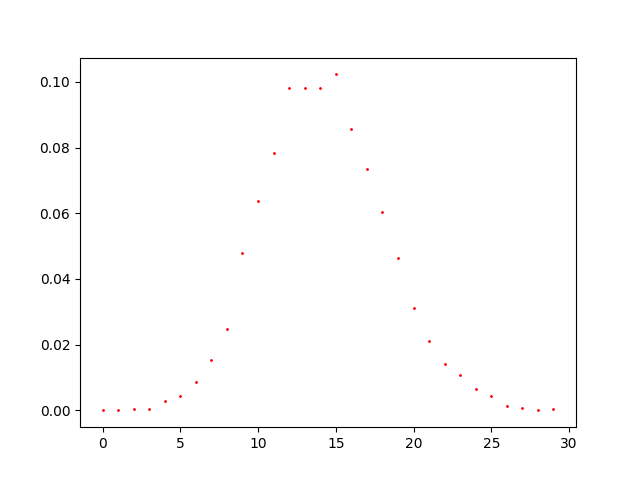
r=1, k=1, q=1



As we can see by the figure, the shape is really similar to the one of the our net. However, it is shifted on left, it picks around 5 and the max degree is 15. So we can tune r and k.

Nodes = 4000

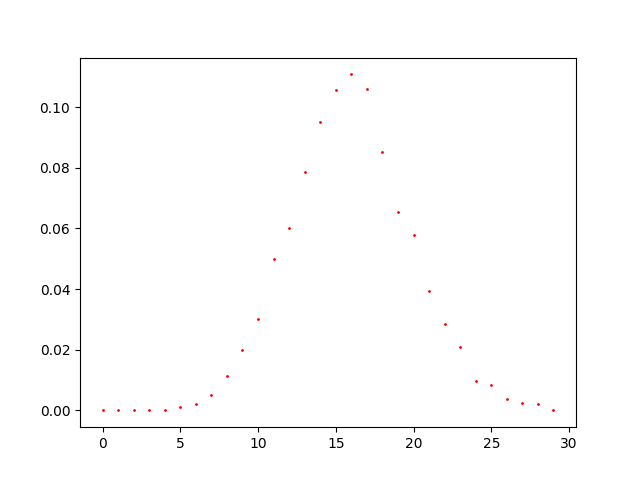
R=2, k=1, q=1



The shape and the degree range is more similar increasing r, but the pick is shifted to left respect the one of the our net. So we should try to increase k.

Nodes = 4000

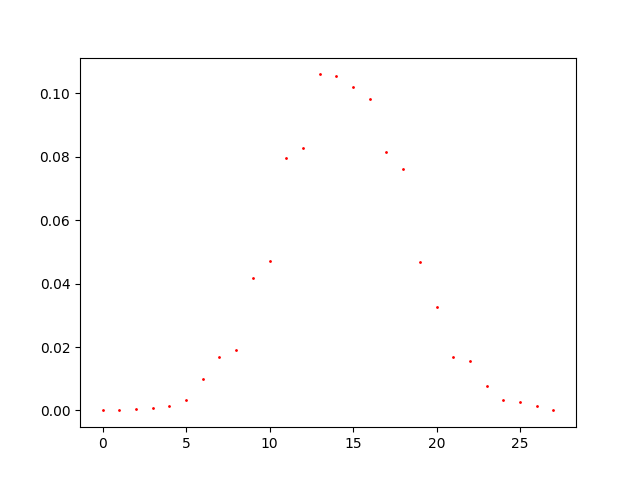
R=2, k=2, q=1



As we can see by the figure, the shape is really like shape of the net 6. However, the range of degrees is slightly greater than our range and the curve starts to stands up around 5, while our curve starts rises after this number. So we want decrease the maximum degree of the network and we want to slightly shift to right the rises of the curve. Now try to increase q:

N=4000

R=2, k=2, q=2

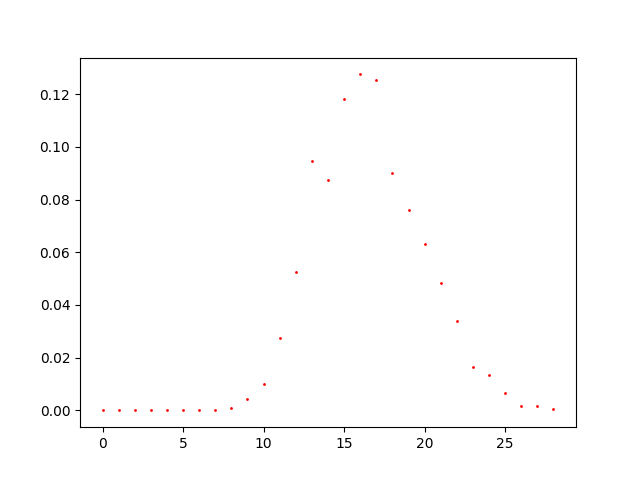


We have decreased the max degree of the net and we have shifted the curve to the left, but now we would shift a little bit to the right side.

After these experiments, we have tried with different configurations. If we want to shift the curve on the right, but we want to keep a slow value for the max range, maybe we could keep the r value to 1 and increase only k. However we choose to decrease the number of nodes to 2000 in order to slow down the time to generate the network.

N=2000

R=1, k=7, q=1



As we can see, we have restricted the degree range, the curve picks at 16 but the curve start rising after 7, while our curve start rising just after 5.

Try to decrease just a little bit the value of k.

N=4000

R=1, k=6, q=1

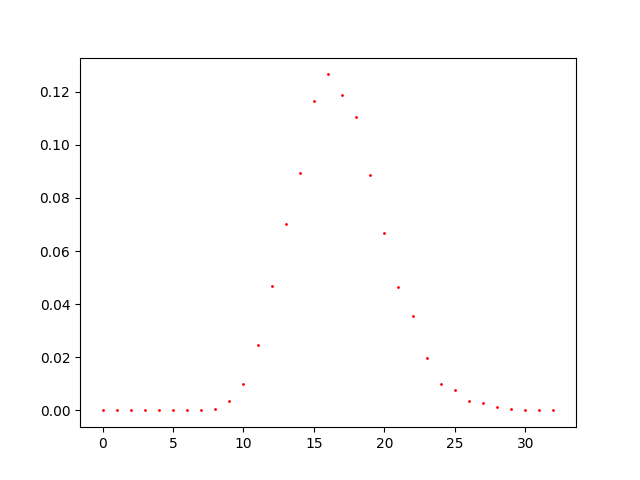


Although these results seems to be near to our curve, with 10000 nodes there are some changes in the behaviours. For this reason we try with 10000 nodes the results we have achieved.

We try with the same values but with 10000 nodes, and the results are not as expected.

N=10000

R=1, k=7, q=1



As we can see, as before, the curve peaks at 16 but it starts rising after 7, while our curve starts rising just after 5. However, the range of degree is greater than before, indeed the max degree before was 29 while now is 33.We try with k=6.

R=1, k=6, q=1

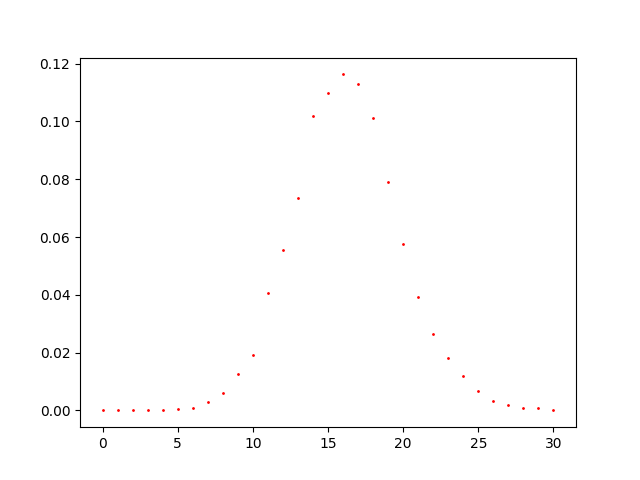


We have a maximum degree of 29, but the pick is shifted on the left.

We can notice the fact that increasing k we shift the pick on the right, but even the range of degrees is shifted. We would shift only the peak, but not the maximum degree value. So we try to increase q, because it is a parameter involved to compute how is probable to have a connection in relation to the distance. However, if we only increase q, then the pick will be shifted on the left, so we should to increase k and r, but r should be increase a little.

We try the following configuration:

R=1.5, k=8, q=2



As we can see by this figure, the rising starts just after 5, the peak is at 16 and the maximum degree is 30. We are so close to our shape that we can conclude that these parameters are those chosen to generate net 6.

Since the ranges of degree could change during the experiments with the same values, in order to obtain the real values we just try a series of experiments using the values we have tried that could be right on a net of 10000. We ran these experiments:

experiments = {  
 1: {"n":10000, "r":1, "k":6, "q":1, "n\_trials":10},  
 2: {"n":10000, "r":1, "k":7, "q":1, "n\_trials":10},  
 3: {"n":10000, "r":1, "k":7, "q":2, "n\_trials":10},  
 4: {"n":10000, "r":2, "k":1, "q":1, "n\_trials":10},  
 5: {"n":10000, "r":1, "k":7, "q":1.5, "n\_trials":10},  
 6: {"n":10000, "r":2, "k":2, "q":1, "n\_trials":9},  
 7: {"n":10000, "r":2, "k":2, "q":1.5, "n\_trials":10},  
 8: {"n":10000, "r":1.5, "k":6, "q":1.5, "n\_trials":5},  
 9: {"n":10000, "r":1.5, "k":7, "q":1, "n\_trials":10},  
 10: {"n":10000, "r":2, "k":2, "q":2, "n\_trials":10},  
 11: {"n":10000, "r":2, "k":5, "q":2, "n\_trials":10},  
 13: {"n":10000, "r":1.5, "k":8, "q":2, "n\_trials":10},  
}

After each trial, we save the figure and the max degree, and at the end the parameters that outputs the same shape for the most times were chosen to show up.

At the end, as we show before, the model we guess is a Watts-Strogatz with those parameters: r=1.5, k=8, q=2.

## Exercise 3

Discussion about exercise 3

### Task

Suppose there is an election and the voters are connected through a social network. G = (V, E). Suppose that there are n voters, represented by the nodes of the graph G, and m candidates. Each candidate i has a position pi in [0,1] that represents her political tendency (for example, a candidate whose position is close to 0 or 1 is, respectively, an extreme-left or an extreme right candidate, while a candidate with position close to 1/2 is moderate).

Each voter u has single-peaked preferences with peak in bu, i.e., she ranks candidates according to the distance of their positions from bu, by breaking possible ties in favour of the candidate on the left of bu (thus, the most preferred candidate is the one whose position is closest to bu, the second most preferred candidate is the one with the second closest position and so on).

The election occurs according to a plurality voting rule (see lesson about voting for a definition). We call an election truthful if each voter u votes for the candidate closest to her peak bu.

On the other hand, a voter can be influenced by opinion campaigns run over the social network and she could be induced to vote a candidate different from her favourite one. Specifically, we consider a manipulator that wants to improve the outcome of a given candidate c. To this aim the manipulator can select at most B voters (in the following called seeds), alter their peaks and use their influence to induce a change in the votes expressed by other voters.

We assume that the voting opinions diffuse over the network according to a FJ dynamic.

Specifically, if S is the set of at most B seeds chosen by the manipulator, then:

1. Set xu(0) = bu, and su=1/2 for every u not in S
2. For u in S, let bu = b’u, where b’u is defined by the manipulator, and set xu(0) = bu, and su=1
3. Run the FJ dynamics with this configuration
4. Once the dynamics reaches the equilibrium at time step t, update the preferences of voters by setting the peak pu = xu(t)
5. Re-run the election with voter’s peaks in pu. We call this election manipulated.

You have to design an algorithm that, given a network G, a set of m candidates with their positions (p1, …, pm), a special candidate c, a budget B, and the initial peaks of all the voters (b1, …, bn), returns a set S of at most B seeds and a peak value b’u for each seed u in S, such that the difference between the number of votes obtained by the candidate c in the manipulated election and the truthful one is maximized.

All the proposed manipulation algorithms will be tested on a common input. The group providing the larger increment in the number of votes for the candidate c will receive a bonus point.

INSTRUCTION FOR THE SUBMISSION:

Your code must include a function manipulation(G, p, c, B, b), where G is an undirected, unweighted graph, p is a Python list with each element in [0,1], c is in {0, …, len(p)-1}, B is a positive integer, and b is a Python list such that len(b) = len(G.nodes()) with each element in [0,1].

The function must print only one string that contains the following three elements separated by a comma:

* the group number;
* the number of votes for candidate c before the manipulation occurs;
* the number of votes for candidate c after the manipulation

### Solution

Proposed solution.