SNOM HW1 - Exercise 2

Michele Luca Puzzo 1783133

May 2022

1 Erdős-Rényi G_{np} random graph model

To build the Erdős-Rényi G_{np} model I have generated an empty graph with n nodes and, through the function *itertools.combination*, a list of all possible edges which could be formed in the graph. They are the all combinations of n nodes taken two-by-two. Then for each possible edge i, I generate a random number between 0 and 1 and if it is lower or equal than p, I add that edge. Indeed in this model edges are independent each other and this is an issue to model a real social network because here there will not be triadic closures and nodes are all the same. Instead in real social networks nodes can be very different very different each other.

	Number of nodes	Prob of an edge to exist	Diameter	Clustering Coefficient
0	10000	0.001	6	0.000939
1	10000	0.010	3	0.010003
2	10000	0.050	2	0.050012
3	10000	0.100	2	0.100094
4	10000	0.300	2	0.299963
5	10000	0.500	2	0.500021
6	20000	0.100	2	0.100003
7	20000	0.010	3	0.009992
8	20000	0.001	5	0.000975
9	40000	0.001	4	0.001004
10	60000	0.001	4	0.001003

Figure 1: Table shows metrics computed with different values of n and p

I have experimented this model for eleven different couples of values of (n,p), in particular at the beginning I have fixed n to 10K and have tried six different values of p from 0.001 up to 0.5. Then I have fixed n to 20K and have tried p equals to 0.01 and 0.1. Finally I have fixed the value of p to 0.001 and I have increased the number of nodes n from 20K up to 60K.

1.1 Diameter

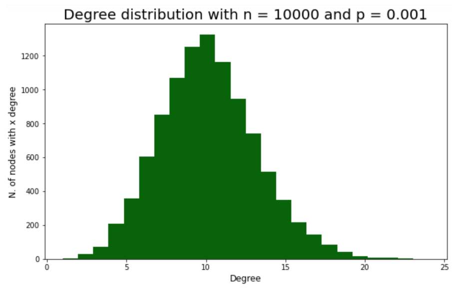
I have always taken $p > \frac{c \ln n}{n}$ $with\ c > 1$ so I was sure that graph is connected[1]. The diameter is the maximum shortest path among all pairs of nodes. Furthermore I knew from [1] for this model it is at most $\frac{2 \ln n}{\ln np}$ so also like the intuition suggests, the less the probability of an edge to exist, the longer the diameter. My experiments show clearly this behaviour: when I have fixed n equals to 20K and decreasing p from 0.1 to 0.001 the diameter moves from 2 to 5. Analogously when I have taken n equals to 10K, but in that for p greater than 0.05 the diameter stays equal to 2. The diameter decreases also when the number of nodes increase and I show this behaviour with the first and last three trials: diameter moves from 6 to 4 when number of nodes moves from 10K to 60K. In the overall, I can confirm that the diameter is small even when the probability p is small, and this is one of the attractive properties of this model. However I will notice later that real social networks have bigger diameters.

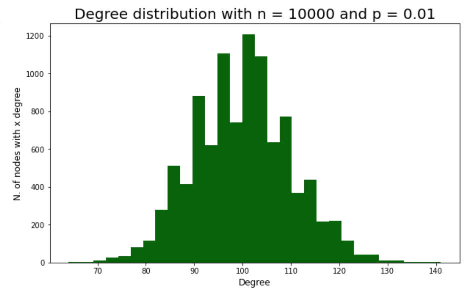
1.2 Clustering Coefficient

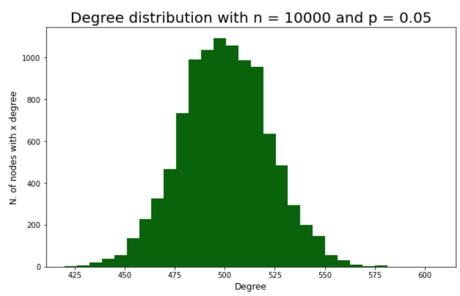
In this model the clustering coefficient C is equal to p, [1], and indeed my trials confirm this notion as it is possible to see in the last column of the table. Since p is usually small the clustering coefficient is also small which is a problem of this model since I will observe later that real social networks are more clustered.

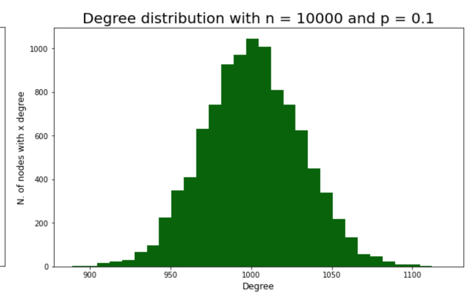
1.3 Degree Distribution

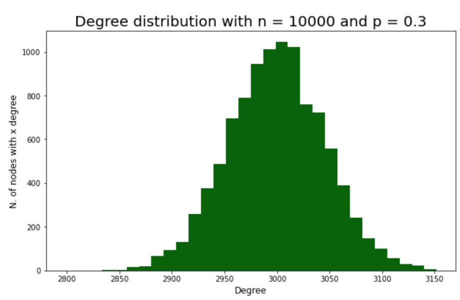
In this model the degree distribution follows a binomial distribution:
$d_V \sim Binomial(n-1, p)$ and so the expected degree is $E[d_V] = (n-1)p \approx np$. Also from the plots of degree distribution it is visible that are binomials that assume their maximum value for $n \cdot p$. From [3] I have read: "If n is large enough, then the skew of the distribution is not too great. In this case a reasonable approximation to B(n, p) is given by the normal distribution $\mathcal{N}(np,\, np(1-p))$. This is an issue to model the real social networks, indeed in them it is observable a heavy-tailed degree distributions because degree distribution follows a Power Law. Just due to the randomicity the bell shape is not in all the simulation completely smooth.

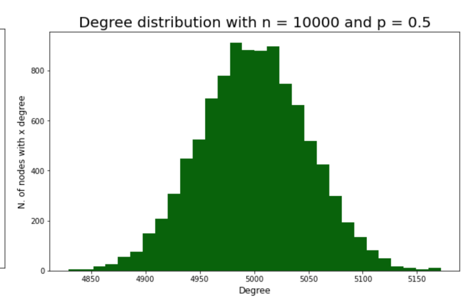

Degree distribution with n = 10000 and p = 0.001
N. of nodes with x degree
0
200
400
600
800
1000
1200
Degree
0
5
10
15
20
25

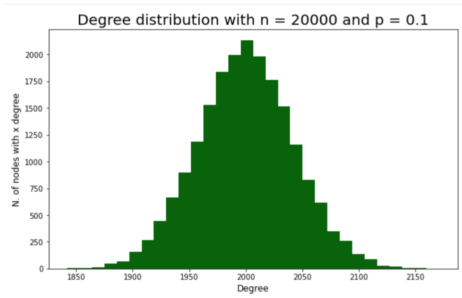

Degree distribution with n = 10000 and p = 0.01
N. of nodes with x degree
0
200
400
600
800
1000
1200
Degree
70
80
90
100
110
120
130
140

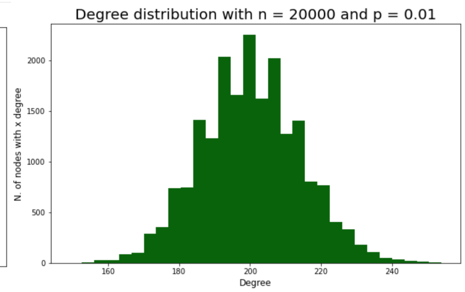

Degree distribution with n = 10000 and p = 0.05
N. of nodes with x degree
0
200
400
600
800
1000
Degree
425
450
475
500
525
550
575
600

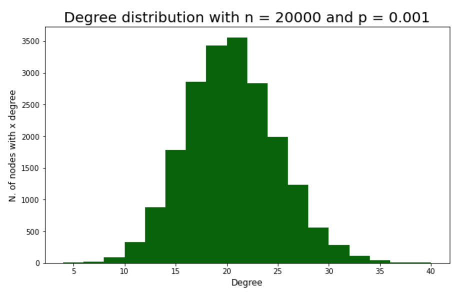

Degree distribution with n = 10000 and p = 0.1
N. of nodes with x degree
0
200
400
600
800
1000
Degree
900
950
1000
1050
1100

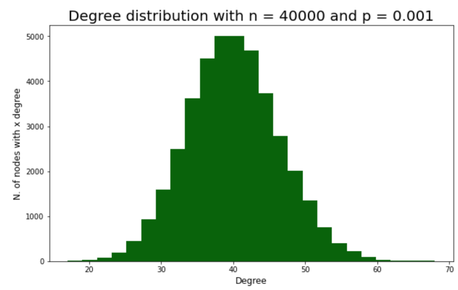

Degree distribution with n = 10000 and p = 0.3
N. of nodes with x degree
0
200
400
600
800
1000
Degree
2800
2850
2900
2950
3000
3050
3100
3150

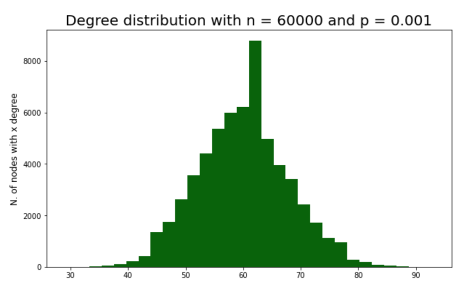

Degree distribution with n = 10000 and p = 0.5
N. of nodes with x degree
0
200
400
600
800
Degree
4850
4900
4950
5000
5050
5100
5150











In conclusion with my trials I have confirmed some of the properties of Erdős-Rényi G_{np} model and highlight its criticisms to model a real social network.

2 Barabási-Albert preferential attachment model

To build Barabási-Albert model I have taken the following decision: "if multiple edges fall on the same node I will ignore the multiple edges". So I have initialized a graph and add the first node at time step t = 0. Due to my initial decision the second node surely form just one edge with the first node at time step t = 1. Then at each time time step t, I have computed the degree d_i for each node i, and have built an array *prob* in which each element i represents the probability of node i to be selected: $\frac{d_i}{\sum_{w \in V_{t-1}} d_w} = \frac{d_u}{2(t-1)l}$ where V_t is the set of nodes at time t. Now through the function *stats.rv_discrete* I have built a discrete distribution probability where the support points are the nodes while the masses that I put on each of them is their degree. Following these probabilities, I pick at most l nodes: since there cannot be multiple edges if I select a node more times, I pick it just one time. Finally I add a node to the graph and build an edge between the new node and the selected nodes.

	Number of nodes	New edges at each step	Diameter	Clustering Coefficient
0	20000	3	7	0.002675
1	20000	5	5	0.003943
2	20000	7	5	0.004723
3	20000	10	5	0.006072
4	20000	15	4	0.007929
5	40000	5	6	0.002208
6	80000	5	6	0.001313
7	100000	5	6	0.001141

Figure 2: Table shows metrics computed with different values of n and l

I have experimented this model for eight different couples of values of (n,l), in particular at the beginning I have fixed n to 20K and have tried five different values of l from 3 up to 15. Then I have fixed l to 5 and I have increased the number of nodes n from 40K up to 100K.

2.1 Clustering Coefficient

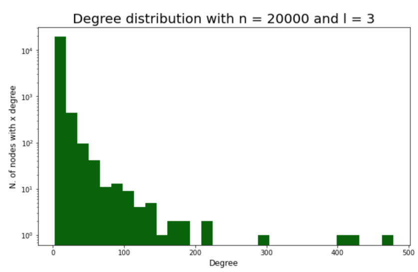
I have given a look at [2] where I have read that clustering coefficient is $\frac{l-1}{8}\frac{(logn)^2}{n}$. The values that I found through with my experiments respect this formula. In general as intuition suggests it increases a little bit when l increases, and it decreases when n increases. However in this model the clustering coefficient is very low so as the random graph model, it does not display the community structure that is quite common in real social networks and that I will observe later.

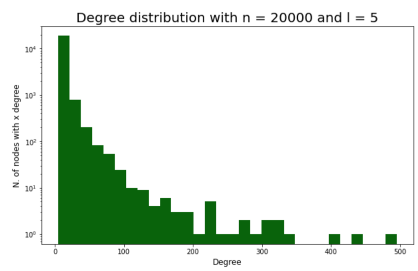
2.2 Diameter

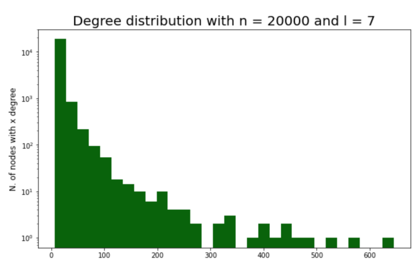
Through the first five experiments I have noticed, as the intuition suggests, that the more new edges at each step the shorter the diameter even if not too much indeed moving from l=5 to l=10 diameter remains constant. On the other hand increasing the number nodes and keeping fix l, also the diameter increases, but much slower indeed from n=40K to n=100K diameter remains equal to 6. So I can conclude that diameter is influenced mainly by l, but also to lesser extent from n.

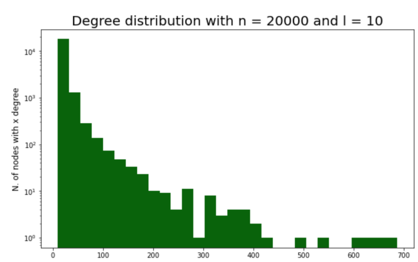
2.3 Degree distribution

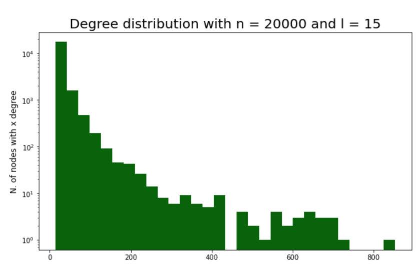
Differently from random graph model where nodes look very similar, in this model *rich-get-richer* phenomenon happens and it leads to power-law degree distributions [1]. It is an heavy tail distribution. The plots of degree distribution that I have obtained from my experiments denote this characteristic of the model. In each plot I can notice that there is always a group of nodes that has a degree much higher than others: probably they are the initial nodes that are the nodes to which new nodes prefer to attach. Then I notice that the greater the l, the heavier the tail.

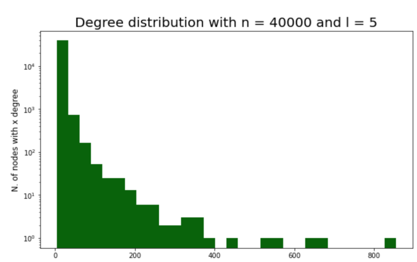


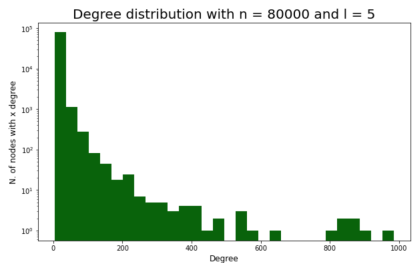


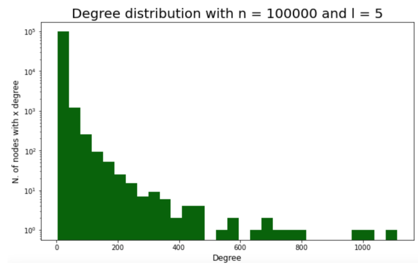












In conclusion with my trials I have confirmed the main characteristic of Barabási-Albert model and highlight its criticism to model a real social network.

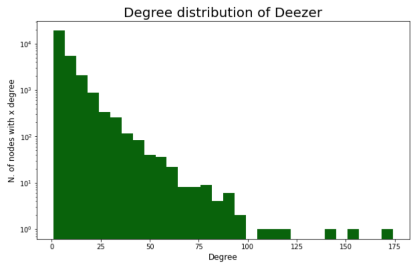
3 Real Networks

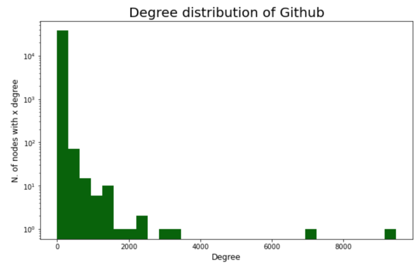
From the SNAP library I have take five graphs that represent social network because they are the network in which I are more interested in. I have chosen undirected graphs which various number of nodes. Computing the clustering coefficient of these real network I noticed that clustering coefficients are higher than values achieved with previous models. As expected, social networks include community groups based on common location, interests, occupation and so on. Instead I have wrongly thought to observe smaller diameters due to the small-world property while instead diameter can assume high values, up to 21. The reason that I give to myself to explain this fact is that there are communities densely connected internally (high clustering coefficient), but there are just few weak ties so communities between them are very far.

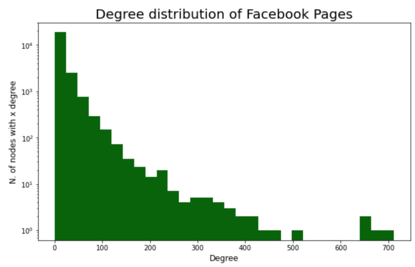
	Network	Number of nodes	Number of edges	Diameter	Clustering Coefficient
0	Deezer	28281	92752	21	0.141160
1	Github	37700	289003	11	0.167537
2	Facebook Pages	22470	171002	15	0.359738
3	Facebook friends	4039	88234	8	0.605547
4	Twitch DE	9498	153138	7	0.200886

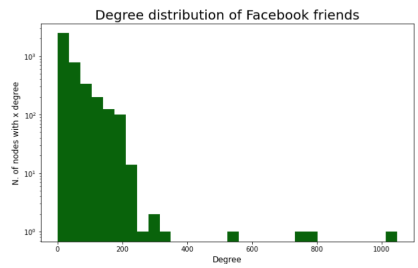
Figure 3: Table shows metrics computed on real social networks

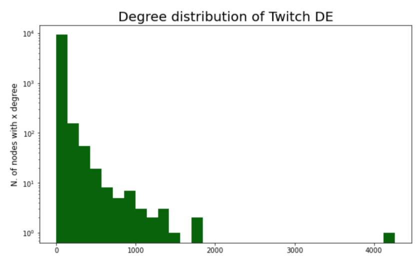
The plots of degree distribution show a power law degree distribution even if obviously the tail of the distribution is different in each plot, for example in Twitch DE networks is heavier than in Github one. These plots are a confirmation that *rich-get-richer* phenomenon happens in real social networks.











4 My model

At the beginning I had built a model combining the random and the preferential attachment model. But in this way I had not obtained sufficiently high clustering coefficients. So I have slightly modified this initial idea. I have started with a graph with n nodes and for each of them I throw an unbalanced coin: with probability p I attach the i-th node to l other nodes chosen uniformly at random, so p is the weight of the Erdős-Rényi random graph model. With probability (1-p) I attach the i-th node to l 'neighbours', the l/2 nodes before and after it. In this way I have wanted to embody the idea that some people prefer being friend with people which are near to them (same school, same environment), while other people prefer being friend with people that do not belong to their environment, quite boring!

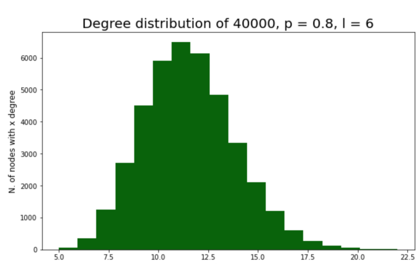
I have experimented this model for seven different triplet of values of (n,l,p), in particular at the beginning I have fixed n to 40K and l to 6, and I have varied the weight of random model p from 0.8 to 0.2. Now keeping again fix n and p to 0.2, I have modified l from 4 to 8. Finally I have fixed again p and l to 8, and I have varied n from 60K to 100K.

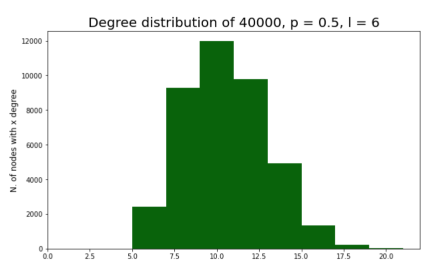
In the first three trials I have noticed that decreasing the weight of randomness the diameter and the clustering coefficient increase. This is reasonable because there are less weak ties, but near nodes are more connected. While keeping fix p, and increasing l the diameter and the clustering coefficient increase a bit because simply there are more edges. Instead if I increase the number of edges keeping constant the other two parameters, the clustering coefficient remains the same while the diameter slowly increases.

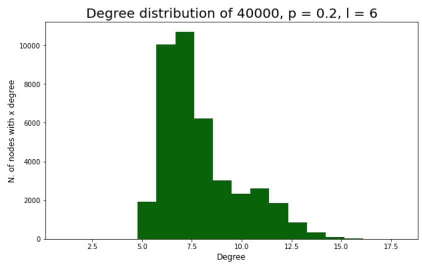
	Number of nodes	New edges at each step	Weight of Erdos model	Diameter	Clustering Coefficient
0	40000	6	0.8	6	0.016591
1	40000	6	0.5	8	0.115949
2	40000	6	0.2	10	0.348328
3	40000	4	0.2	13	0.288290
4	40000	8	0.2	8	0.366304
5	60000	8	0.2	8	0.367880
6	100000	8	0.2	9	0.365284

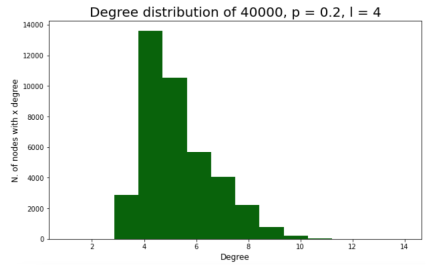
Figure 4: Table shows metrics computed on real social networks

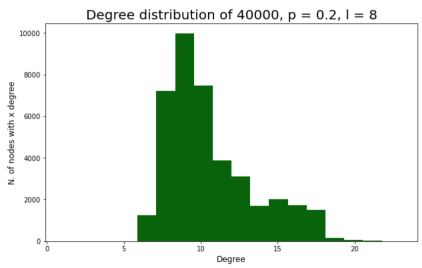
The plots of the degree distribution strongly depend on the weight of the random graph model indeed if p is greater or equal to 0.5 the degree distribution is rightly like a binomial that we have already observed in the first section but here the maximum it seems that is reached around $l + l \cdot p$. On the other hand for low values of p the degree distribution is not more symmetric, indeed all the nodes have at least degree l and the maximum is immediately after and then the histogram falls down quickly because the majority of the nodes have degree around l.

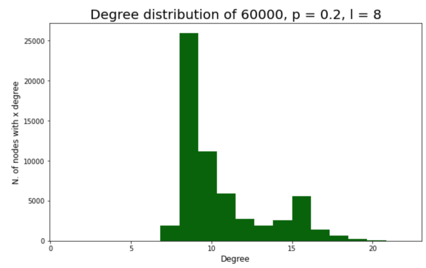


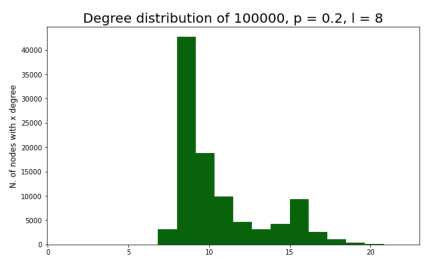












In conclusion through my model I have obtained clustering coefficient comparable with ones of the real networks thanks to mechanism of connect nodes with their 'neighbours'. However the graph is connected and diameter is not too high thanks to the contribution of the random model. The shortcoming of my model is that the degree distribution is completely unrealistic indeed it is not credible that a person has just l, or a bit more, friends. Moreover in real social networks are present hubs, and the degree distribution follows a power law and degree of various people is not quite homogeneous like in my model. On the other hand I am happy to have achieved similar values of diameter and cluste

References

[1] Anagnostopoulos Aris, *"Social Networks and Online Markets – Notes"*.

[2] Fronczak Agata et al., *"Mean-field theory for clustering coefficients in Barab ́asi-Albert networks"*, arXiv:cond-mat/0306255.

[3] Wikipedia, *https://en.wikipedia.org/wiki/Binomial_distribution*