CSN 05 - Network Dynamics

Laura Cebollero, Pietro Fronte

3rd of December, 2018

Introduction

In this delivery we are asked to simulate three network growth models and study them from a statistical point of view.

The first one is the original Barabasi Albert model, and the other two are based on this one with some variations:

1. BA: growth & preferential attachment
2. BA’: growth & random attachment
3. BA’’: No growth & preferential attachment

We are going to discuss the results and the methods used, as well as cover the implementation done and problems encountered.

Implementation

As asked, we have implemented the simulation on these three models with a fixed n_0, m_0 and t_{max}, their values being:

- $n_0 = 3$
- $m_0 = 2$
- $t_{max} = 10.000$

This means that for the models with growth (B.A. and B.A.’), on $t = 0$, there will exist 3 nodes fully connected. Meaning we are starting with a fully connected graph. This decision is taken in order to avoid ending up with a disconnected graph, since if we start with a disconnected graph, after the first iteration has been performed, the non-selected node will never be selected and thus we will end up with a disconnected node from the overall graph.

Each time t increases on every iteration, a stub is going to be added. Each stub will consist on a node with m_0 edges. In our case, $m_0 = 2$, so each iteration we are going to add two edges to the graph.

This will be performed until reaching t_{max}, when the total number of nodes N will be

$$N = n_0 + t_{max} = 3 + 10.000 = 10.003$$

In the third case, B.A.’’ (No growth + preferential attachment), we start with 2.000 nodes and end with the same number of nodes. We have chosen a somewhat big number of nodes as suggested in the lab session statement. More specifically, the suggestion was that $N \geq 1000$, and t_{max} has been preserved as in the other two models to preserve coherency. Having a greater t_{max} would have led to a complete graph which is not what we want to achieve.

In those 2 cases where preferential attachment is what determines where each edge from the stub connects to, the higher the number of edges an existing node has, the higher the probability it has to be the one selected for the new addition.

In the first model and second model we do not have to worry on whether we are creating multiedges or not, for each stub is added in every iteration and it cannot select itself. Because of this, we have not used an adjacency matrix and only have tracked the degree evolution of 4 selected nodes, as well as the global degree sequence when reaching t_{max}.

However, to avoid having multi-edges on the third model, we have used a sparse adjacency matrix to check whether an edge existed or not before adding it or not. The sparse decision has been taken in order to be more lenient on our computer memories. The usage of such matrix makes the execution of this simulation a little bit slower compared to the other ones.

Below is a table that represents on average the time it has taken us to execute the three simulations:

Elapsed times on simulation runs

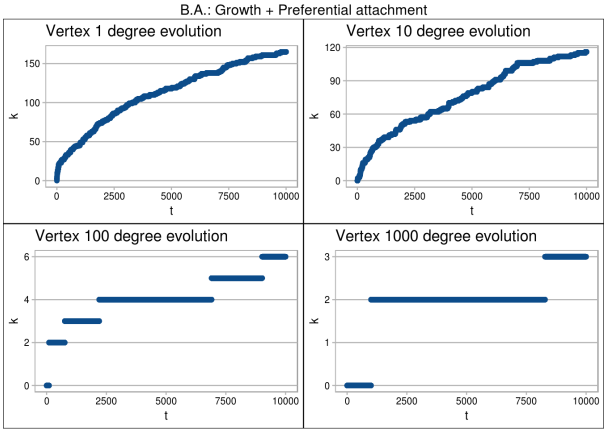
Simulation	$\Delta t(seconds)$
B.A.	5
B.A.’	3
B.A.’’	35

As mentioned, we can see how the third simulation (No growth + preferential attachment) is the one which has more cost computationally speaking because of the usage of the adjacency matrix.

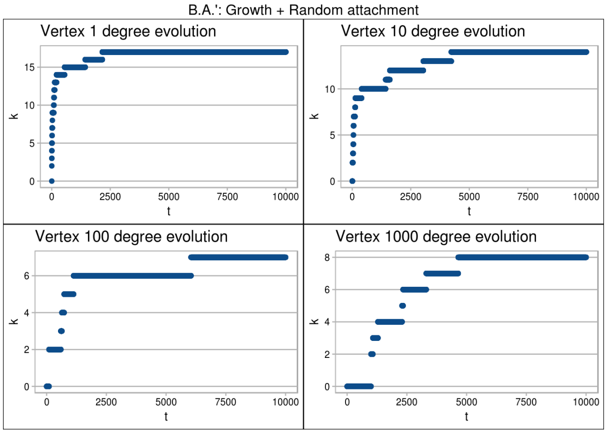
Results

Now we are going to see how is the evolution of 4 vertices until reaching t_{max}, as well as what distribution better fits each model.

Evolution



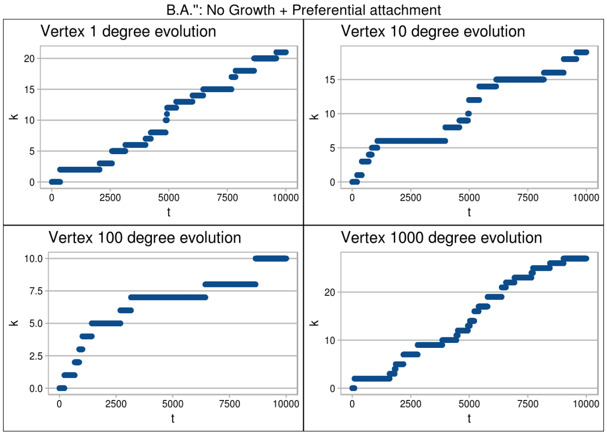
In the plot above we can observe how the first vertex is selected a lot because of the preferential attachment. As the vertex selected chosen is that from a stub added forward in time, clearly its edge degree is going to be lower, thus having a lower preferential attachment for the next iterations. Thus, vertex 100 reaches its highest point at 6 edges, and edge 1000 just 3: the ones created when the node was added as a stub and another iteration where it got selected.



Now that the random attachment is done, we can see how the vertices have more or less the same degree distribution. However, since it is a growth model the first vertices (i.e. 1 and 10) still have more chances to be selected on the first iterations, since the number of existing nodes is lower.

That’s why their edge degree reach a higher point than vertices 100 and 1000.

We can also see how on the last 9000 iterations, vertex 1000 got luckier than vertex 100, for it has one more edge than the other one. Thus, we can see how the randomness plays its hand on the attachment of stubs.



In this case we can see how the no-growth and preferential attachment have their effects on the simulation model.

On the first iteration all vertices (2000) have the same chance to get selected. Afterwards, those that got lucky and got selected on the first iterations have a higher probability to get selected again.

In this case, it seems that vertex 1000 got lucky and got selected a lot more than its 3 other counterparts.

Scaling of vertex degree over time

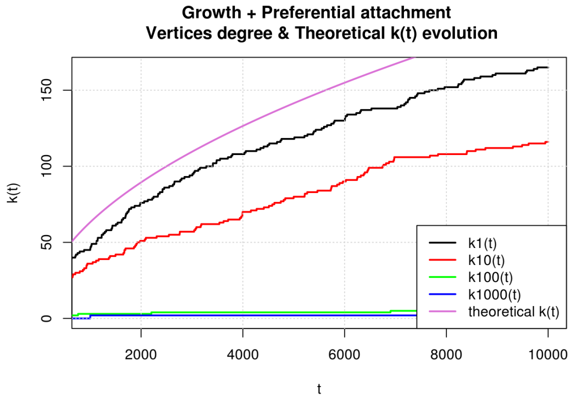
Now let’s take a look on how these vertices’ evolution grow comparing them with the theoretical k_i. Each simulation has its own theoretical average growth.

Growth and preferential attachment

On the first simulation model, the formula to follow is

$$k_i''(t) = m_0^{t_0} = m_0\sqrt{t_0}$$

This theoretical evolution has been painted in pink in the plot below.



We can see how vertex 1 follows somewhat closely that theoretical evolution. Of course, the vertices added later in time do not follow this evolution, for they have a lower chance to get chosen on the preferential attachment. The plot focus only in the range (1000,t.max), range where all the 4 selected points coexist.

	Model	RSS	AIC	Param1	Param2	Param3
ks.1	log+i	36050.930	41210.25	82.1391883	1887.4933536	-604.8392698
ks.2	plaw+i	81148.158	49323.69	0.5998824	0.5750098	0.5271861
ks.3	plaw+i	1693.116	10626.63	0.1610148	0.3702034	0.5362697
ks.4	log+i	1770.000	11070.71	0.5345786	-377.4945910	-2.3313333

In this part of the assignement we are asked to verify if the distribution that better fits the scaling of selected vertices is a power-law distribution with gamma $\gamma = \frac{1}{2}$. In order to verify it we perform a model selectio task as suggested in the statement. The results found are the following:

- Scaling of vertex 1, for our model selection process, follows a logarithmic distribution. In details:

$$k_1(t) \sim a\log(t + d_1) + d_2 = 82.14\log(t + 1887.5) - 605$$

- Scaling of vertex 10, as previously expected, follows a power law distribution with a slightly modified $$ parameter. Instead of 0.5 tend more to 0.58

$$k_{10}(t) \sim at^b + d = 0.6t^{0.58} + 0.53$$

- Scaling of vertex 100, as for vertex 10, follows a p-law distribution but with a more free exponend.

$$k_{100}(t) \sim at^b + d = 0.16 + t^{0.37} + 0.54$$

- Scaling of vertex 1000 appears really flat and it is diffult to get an idea of it visually. However the result obtained by model selection is the following:

$$k_{1000}(t) \sim a\log(t + d_1) + d_2 = 0.53\log(t - 377.5) - 2.33$$

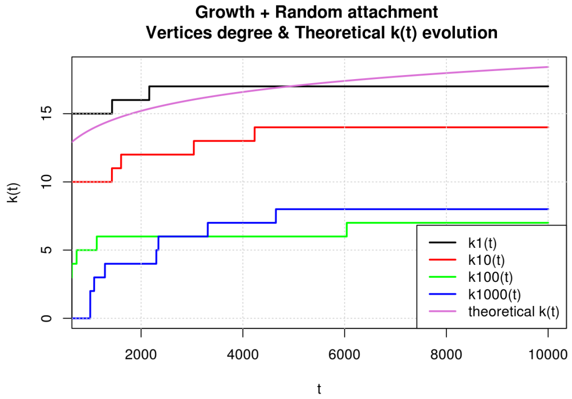
  Notice that, since vertex exists only from t=1000 on, the argument of the logarithm will be always positive.

Growth and random attachment

Now the scalingof vertex degree over time is as follows:

$$k_i''(t) = m_0 log(m_0 + t - 1)$$

We can see how the first vertex follows it closely and the other 3 also tend to stagnate (later, of course, since they are added later on).



	Model	RSS	AIC	Param1	Param2	Param3
ks.1	plaw+i	3144.762	16818.30	101.344332	0.0099873	-93.50548
ks.2	plaw+i	2261.748	13522.30	44.651677	0.0300751	-44.31388
ks.3	plaw+i	2783.591	15598.34	44.860469	0.0222808	-47.81551
ks.4	log+i	6050.741	23362.73	2.997463	235.7301761	-18.64355

When replacing preferential attachment with random attachment we are modifying the growth of the degree of each vertex. In this case all of them, in the range where they exist, are equally probable. With random attachment the theoretical scale of vertex should follow a logarithmic distribution and we are asked to check if actually this distribution is the one, with appropriate parameters, that better approximate the data collected.

- Scaling of vertex 1, p-law with intercept as best distribution selected. In details:

$$k_1(t) \sim at^b + d = 101.34t^{1e^{-3}} - 94$$

- Scaling of vertex 10, p-law with intercept as best distribution selected.

$$k_{10}(t) \sim at^b + d = 44.6t^{0.03} - 44.3$$

- Scaling of vertex 100, again p-law with intercept.

$$k_{100}(t) \sim at^b + d = 44.8t^{0.02} - 47.8$$

- Scaling of vertex 1000 is the only one that the model selection process find to be better approximated by a logarithmic distribution.

$$k_{1000}(t) \sim a\log(t + d_1) + d_2 = 3\log(t + 235.7) - 18.6$$

We can also see, as suggested by the statement, that a ≈ m0 = 2. We can’t say however the same for d_1, expected to be around m0-1 and obtained around 235.

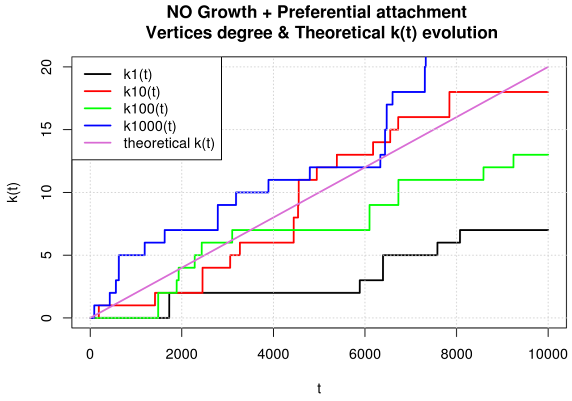
No Growth and preferential attachment

Now

$$k_i(t) = \frac{2m_0}{n_0}t$$

for every vertex.

We can see how, more or less, they all follow this vertex degree scaling. They vary a little because of the preferential attachment.



	Model	RSS	AIC	Param1	Param2	Param3
ks.1	plaw+i	4858.478	21168.17	0.0000023	1.632635	0.2730452
ks.2	log+i	17195.919	33807.64	72.3088123	27725.655000	-741.5176832
ks.3	log+i	9017.909	27353.04	11.1488000	3904.067668	-93.5483023
ks.4	plaw+i	35608.951	41086.89	0.0000028	1.751934	4.0602071

In presence of No Growth the theoretical scaling we are expecting to find is a linear model. However the results obtained don’t confirm this idea. These are the results found:

- Scaling of vertex 1 follows a power low distribution with these parameters:

$$k_1(t) \sim at^b + d = 2e^{-7}t^{1.63} + 0.27$$

-Scaling of vertex 10 follows a logarithmic distribution with intercept:

$$k_{10}(t) \sim a\log(t + d1) + d2 = 72.3\log(t + 27725.6) - 741.5$$

-Scaling of vertex 100 approximated again with a logarithmic distribution with intercept:

$$k_{100}(t) \sim a\log(t + d1) + d2 = 11\log(t + 3904) - 93.5$$

-Scaling of vertex 1000, this time vertex with the highest degree among the 4 nodes, approximated with p-law distribution:

$$k_{1000}(t) \sim at^b + d = 2e^{-7}t^{1.75} + 4$$

Degree distribution

Now let’s check which final degree distribution fits better each simulation model.

To select the best distribution fitting our final degree sequence of each model, we are going to use the functions used on lab. session 2.

Afterwards, we are going to select the model that has the lowest AIC value.

Growth and preferential attachment

First we compute the log likelihood of each distribution fitting our degree sequence.

The table below summarizes the AIC and best parameter obtained from each distribution.

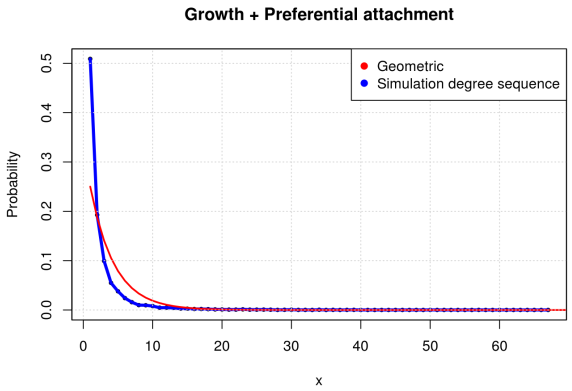
Model	X2LL	AIC	Param1	Param2
Poisson	63152.68	63154.68	3.9197327	NA
Geometric	44995.13	44997.13	0.2500562	NA
Zeta	52262.77	52264.77	1.6275755	NA
RT Zeta	77291.64	77293.64	1.9009266	243

In this case, the distribution selected that has the lowest AIC, thus the model that best fits the final degree sequence is:

```
## Best fitting for B.A. Growth + preferential: Geometric  with parameter value 0.250
0562
```

In the table above we can see there is a large margin between the Geometric distribution’s AIC and the rest.

Let’s see how this geometric distribution fits our degree distribution:



As we can see above, the selected geometric distribution fits quite well our simulation degree sequence.

Growth and random attachment

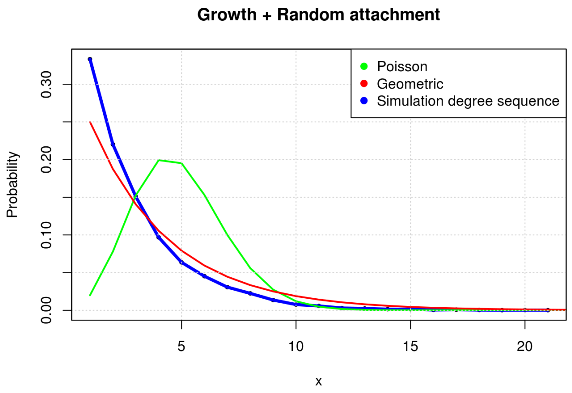
Now we are going to compute the log likelihood of each distribution fitting our degree sequence again for this new degree sequence:

Model	X2LL	AIC	Param1	Param2
Poisson	43145.36	43147.36	3.919413	NA
Geometric	44993.40	44995.40	0.250075	NA
Zeta	56335.76	56337.76	1.578978	NA
RT Zeta	80346.74	80348.74	1.712222	24

Thus, on the second model, the best fitting distribution is:

```
## Best fitting for B.A. Growth + random attachment: Poisson  with parameter value 3.
919413
```

We can see that the selected Poisson distribution is closely followed by the geometric function, so we are going to plot both of them to see how they fit our true degree sequence distribution:



We can see that although the selected distribution that had the lowest AIC is the Poisson distribution. However, visually speaking, we can see that maybe it is the Geometric one that better fits our data.

We have not found the reason for this miss-selection.

TODO: Piero, maybe take a look at this plot above too? :/

No Growth and preferential attachment

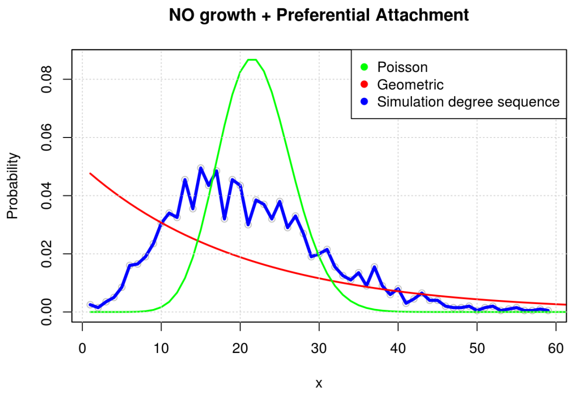
Finally, on the third model, let's see what the best distribution fitting our sequence is:

Model	X2LL	AIC	Param1	Param2
Poisson	18592.28	18594.28	21.000000	NA
Geometric	16081.30	16083.30	0.047619	NA
Zeta	20674.17	20676.17	1.289861	NA
RT Zeta	27011.82	27013.82	1.432330	76

We can see that the Geometric distribution is the one with the lowest AIC, followed somewhat closely by the geometric distribution.

```
## Best fitting for B.A. Growth + random attachment: Geometric  with parameter value
0.04761905
```

Let's plot the degree sequence against the poisson distribution and geometric one so we can compare them:



We can see how both the Poisson and the Geometric distribution may somewhat our simulation degree sequence. However, the obtained λ is very large (21), which is what may affect this big peak around x = 20, thus making the Geometric distribution a better candidate.

Conclusions

To finish with this project, we are going to summarize our conclusions on both the study on vertex scaling in the three models as well as their degree distributions study.

Vertex scaling

In all the three models we have seen, graphically, how they follow the theoretical vertex scaling pretty closely.

For the first two cases, since they are growing methods we can see how the evolution of the first vertex follows the theoritical one.

However on nodes that have been added afterwards (i.e. 100) this scaling does not apply on the first model, or the growing stagnates earlier on the second one because of the preferential attachment.

And finally, in the third model, since it is a no growth model, all the vertices seem to follow more or less the theoretical scaling vertex.

From a model selection point of view we had some results that didn’t match our expectation and what we were asked to find from the statement. We don’t even know where to locate a possible bug: either in the generating process of the graph or in the selection model task to pick the correct model. However, except for the No Growth + Preferential where a linear model is expected if looking at the plot and then not obtained, the result have been matching, overall, what found graphically.

Degree distribution

We have seen how on the first Barabasi Albert model (Growth + preferential attachment), the final degree distribution clearly follows a Geometric distribution.

However, on the other two models it is unclear to us whether the selected distributions are correct or not.

In the second model (Growth + random attachment), Poisson is selected because it has a lower AIC. Although if we take a look at the plot, visually speaking the Geometric distribution seems to fit better our degree sequence.

And in the third model (No growth + Preferential attachment), the Geometric distribution is selected although Poisson could have been selected if the lambda were to be lower. We can see how it follows a somewhat Gaussian distribution with a head and tail on probability 0, which the Geometric does not have (it only has a decreasing tail).

So, to summarize, our conclusions are that there does not exist a clear degree distribution that best fits our degree sequences on the last two models.

Future work

Some possible future work would be to see whether the best fitting distributions change when

$$n_0 \geq 3$$

and

$$m_0 \geq 2$$

Since we have not had enough time to perform these tests, we have made these two parameters static instead of dynamic. It would be interesting to see the effects on changing them, although it may probably be very time