Network dynamics

Marta Arias, Ramon Ferrer-i-Cancho, Argimiro Arratia

Version 0.5
Complex and Social Networks (2018-2019)
Master in Innovation and Research in Informatics (MIRI)

1 Introduction

In this session, we are going to simulate different network growth models and analyse their properties from a statistical perspective. This session has three goals: achieving a better understanding of the dynamical principles behind the Barabási-Albert model, improving your simulation skills, and applying curve fitting methods (model selection) from lab session 2. As you know, the Barabási-Albert model is based on two dynamical principles: vertex growth (at every time step one vertex is added) and preferential attachment (the new vertex is connected to existing nodes with a probability proportional to the degree of the target node).

We borrow the notation from the theoretical sessions; t indicates the time step ($t \geq 0$) and t_i refers to the time at which the i-th vertex appeared. The model has two parameters

- n_0: the initial number of vertices.
- m_0: the initial number of edges of every new vertex.

Besides these parameters, the initial set of edges can also influence the dynamics of the model.

This session consists of simulating and analyzing the mathematical properties of the Barabási-Albert model and two modified versions: one where preferential attachment is replaced by random attachment and another where vertex growth is suppressed [Barabási et al., 1999]. For every variant you will have to study the growth of vertex degrees over time and their degree distribution.

1.1 Simulation

You will have to simulate every version of the Barabási-Albert model till time t_{max} and produce two kinds of output for each variant:

- A degree sequence: a file with the degree sequence at time t_{max}
- Times series: four files with the evolution of the degree of different vertices as a function of time. You have to choose four different arrival times and trace the evolution of the degree of vertices arriving at each of those times. We suggest that the arrival times of the vertices chosen are 1, 10, 100 and 1000 (do not use vertices arriving at time 0).

The time t_{max} should be of the order of 10^4 or 10^5.

For simulating the Barabási-Albert model and its variants, you do not need to use `R`. However, we strongly recommend using `R` for model selection.

1.2 Model selection

For the analysis of the degree distribution, you have to apply the methodology presented in lab session 2. The only difference is that you have to change the definition of model 3: the $\gamma = 2$ exponent of the zeta distribution has to be replaced by $\gamma = 3$, the exponent produced by the standard Barabási-Albert model.

For the analysis of the growth of vertex degrees over time, you will have to apply non-linear regression techniques to fit the degree of a vertex as a function of time. You may consider using `R`'s `nls(...)` function to fit models' parameters. Finally, to perform model selection you may use the fact that for a non-linear regression model, the AIC can be calculated with [Ritz and Streibig, 2008, p. 105; Eq. 7.2]

```
AIC <- n*log(2*pi) + n*log(RSS/n) + n + 2*(p + 1),
```

where p is the number of parameters of the model, n is the number of points in the data, and RSS is the residual sum of squares (sum of squared differences between model's prediction and observed value).

For $f(t)$, the degree of a vertex at time t, you will have to consider the following ensemble of models:

- Model 0: $f(t) = at$ (linear model with 0 intercept)
- Model 1: $f(t) = at^{1/2}$ (expected growth in the standard Barabási-Albert model)

- Model 2: $f(t) = at^b$ (generalized power-law growth)
- Model 3: $f(t) = ae^{ct}$ (exponential growth)
- Model 4: $f(t) = a\log(t + d_1)$

and their variants with an additive term

- Model 0+: $f(t) = at + d$ (full linear model)
- Model 1+: $f(t) = at^{1/2} + d$
- Model 2+: $f(t) = at^b + d$
- Model 3+: $f(t) = ae^{ct} + d$
- Model 4+: $f(t) = a\log(t + d_1) + d_2$

2 The Barabási-Albert model: growth + preferential attachment

2.1 Simulation

To simulate the original version of the model, it is convenient to use a vector of stubs (an edge connecting vertices u and v gives two stubs, i.e. u and v). The number of position of this vector that are occupied at time t before adding the new vertex is $s_0 + 2m_0(t-1)$, where s_0 is the initial number of stubs (the number of stubs added at time 0), and $2m_0(t-1)$ is the number of stubs added before time t. A new vertex arrives at time t and then stubs are chosen at random till m_0 edges can be formed (the same vertex cannot be chosen more than once; multiedges are not allowed). It is easy to show that the probability that a certain vertex is chosen is proportional to its degree, that is satisfying the preferential attachment rule. Notice that the current degree of a vertex is equivalent to its number of stubs in the vector.

- To generate a uniform random integer, do not use `rand()` (in C/C++). It produces small random integers.
- We recommend the `random` library from the Standard Template Library of C++. See `http://www.cplusplus.com/reference/random/uniform_int_distribution/`.

2.2 Scaling of vertex degree over time

The growth of k_i the degree of the i-th vertex as a function of time obeys

$$k_i(t) \approx m_0 \left(\frac{t}{t_i}\right)^{1/2}.$$

This means that

$$\begin{aligned} k'_i(t) &= t_i^{1/2} k_i(t) && (1) \\ &\approx m_0 t^{1/2} && (2) \end{aligned}$$

should be about the same for every vertex, regardless of its arrival time.

> $k'_i(t)$ indicates a rescaled variant of $k_i(t)$; it **does not** indicate 1st derivative of $k_i(t)$.

Exercise: take the files with the evolution of degree as function of time and

- Check visually if $k'_i(t)$ is about the same for every vertex chosen for the ranges of time the vertices coexist (make a plot).
- Check if the power-law dependency with 1/2 exponent gives the best fit to all the time series. Use model selection by (1) using non-linear regression to estimate best parameters for each of the functions listed in Section 1.2, and (2) using AIC as indicated in Section 1.2 to select the best model.
- Add the theoretical power-law defined by Eq. 2 to the plot.

> - Restrict the analysis of evolution $k'_i(t)$ to the interval of time between t_i and t_{max}.
> - Do not try to complete the curve by assuming, for instance, that $k'_i(t) = 0$ for $t < t_i$.
> - Apply these ideas to the analysis of variants of the Barabási-Albert model.

2.3 Degree distribution

We know that the probability that a vertex has degree k is [Caldarelli, 2007, Barabási et al., 1999]

$$p(k) \approx ck^{-\gamma} \quad (3)$$

with $\gamma = 3$ and

$$c = \frac{2m_0^2 t}{n_0 + t}. \tag{4}$$

Exercise:

- Check if the distribution giving the best fit is a power-law with a -3 exponent (modelled as a zeta distribution or a right-truncated zeta distribution, both with a -3 exponent). Use model selection (lab session 2).
- Plot the distribution giving the best fit and the empirical distribution.

- Notice that the definition of $p(k)$ in Eq. 3 is not suitable for model selection because it is approximated. The most important limitation is that the factor c does not warrant that

  $$\sum_{k=1}^{k_{max}} p(k) = 1.$$

  Thus, the definition of $p(k)$ in Eqs. 3 and 4 should not added to the ensemble of functions.
- Tentatively, the definition of the probability in Eqs. 3 and 4 can be replaced by a zeta distribution with $\gamma = 3$ or a right-truncated zeta distribution. The dependence of the maximum degree on time (the degree of a node cannot exceed the number of vertices - 1) along with the time dependence of c (Eq. 4) suggest that the right-truncated zeta is more appropriate but we want to let standard model selection methods indicate which one is the best after all.

3 Growth + random attachment

3.1 Scaling of vertex degree over time

If preferential attachment is replaced by random attachment, the growth of k_i (the degree of the i-th vertex) as a function of time obeys [Barabási et al., 1999]

$$k_i(t) \approx m_0(\log(m_0 + t - 1) - \log(n_0 + t_i - 1) + 1)$$

This means that

$$\begin{aligned} k_i''(t) &= k_i(t) + m_0 \log(n_0 + t_i - 1) - m_0 \\ &\approx m_0 \log(m_0 + t - 1). \end{aligned} \tag{5}$$

should be about the same for every vertex, regardless of its arrival time.

$k_i''(t)$ indicates a rescaled variant of $k_i(t)$; it **does not** indicate the 2nd derivative of $k_i(t)$.

Exercise: take the files with the evolution of degree as function of time, obtain $k_i''(t)$ and

- Check visually if $k_i''(t)$ is about the same for every vertex chosen for the ranges of time the vertices coexist (make a plot).
- Check if Eq. 5 holds through the logarithmic function of model 4 (or 4+). Check if that model gives the best fit to all the time series (use model selection). If that is the case check that $a \approx m_0$ and $d_1 \approx m_0 - 1$.
- Add the theoretical logarithmic function defined by Eq. 5 to the plot.

3.2 Degree distribution

When preferential attachment is replaced by random attachment, the probability that a vertex has degree k is [Barabási et al., 1999]

$$p(k) \approx Be^{-\beta k}$$

with $B = e/m_0$ and $\beta = 1/m_0$. This approximate functional dependency can be modelled by means of a displaced geometric distribution i.e.

$$p(k) = \frac{\pi}{1-\pi}(1-\pi)^k \tag{6}$$

with $\pi = 1 - e^{-\beta}$ (see the Appendix). We suggest considering a displaced geometric distribution. However, a right-truncated geometric distribution might be a better model (let us know if you think the simple displaced geometric distribution does not suffice).

Exercise 1:

- Check that the distribution giving the best fit is no longer a power-law. The function giving the best fit should be a geometric distribution or a variant (see the Appendix).

Exercise 2:

- Compare the quality of the fit of the best function with that of a geometric distribution.
- Plot the distribution giving the best fit and the empirical distribution.
- If the function giving the best fit is the geometric distribution, check that the value of π giving the best fit satisfies $\pi \approx 1 - e^{-1/m_0}$.

4 No growth + preferential attachment

4.1 Simulation

The network has n_0 vertices (their number does not change over time in this version of the model). At every time step, choose a vertex at random and add m_0 connections to it following the preferentiall attachment rule. Make sure that n_0 is large ($n_0 \geq 1000$). Some of the theoretical results employed below need n_0 large so as to provide a good approximation.

4.2 Scaling of vertex degree over time

If growth is removed, the growth of k_i the degree of the i-th vertex as a function of time obeys (for large n_0 and large t; $t \geq n_0$ is required) [Barabási et al., 1999]

$$k_i(t) \approx \frac{2m_0}{n_0} t. \tag{7}$$

To see it, notice that the number of stubs at time t is $2m_0t$ and then Eq. 7 coincides with the expected number of stubs that belong to the i-th vertex at time t under the assumption of independence. As all vertices arrive at the same time, Eq. 7 indicates that all vertices grow approximately in the same fashion (on average) for sufficiently large t.

Exercise: take the files with the evolution of degree as function of time, obtain $k_i(t)$ and

- Check visually if $k_i(t)$ is about the same for every vertex chosen (make a plot).
- Check if Eq. 7 holds through model 0 (or 0+). Check if model 0 (or 0+) gives the best fit to all the time series (use model selection). You may have to exclude points for $t < n_0$).
- Add the theoretical curve defined by Eq. 7 to the plot.

4.3 Degree distribution

When vertex growth is removed, the degree distribution evolves over time. Initially it is power-law, then Gausian-like and finally a Kronecker delta function [Barabási et al., 1999]. Notice that the final state is a complete graph, i.e. all vertices having degree n_0-1. For sufficiently large t (in the intermediate regime), the distribution changes and vertices tend to have the same degree. Recall Eq. 7. Then, it is expected that degrees are distributed around $\frac{2m_0}{n_0} t_{max}$. This suggests that the degree distribution should be closer to a binomial, a distribution that was not considered in the ensemble of distribution of lab session

2. However, you do not need to add that distribution to that ensemble. It is known that the Poisson distribution with parameter λ approximates a binomial distribution of parameters N and π when $\lambda = Np$. The approximation is good if $N \geq 100$ and $Np \leq 10$ [1]. Thus, do not use small values of n_0 and use a sufficiently large value of t_{max} (a very large value of t_{max} can be problematic because the model converges to a complete graph, where only one degree has non-zero probability; the distribution is trivial in that case).

Exercise:

- Check that the distribution giving the best fit is no longer a power-law for sufficiently large t. The function giving the best fit should be Gausian-like.
- Plot the distribution giving the best fit and the empirical distribution.

5 Deliverables

You have to prepare a report including the following sections (in this order): introduction, results, discussion and methods.

Your report should indicate in the right place your choices for the values of n_0, m_0, t_{max} or the initial configuration of the network at time 0 (this is critical for the variants of the Barabási-Albert model where the preferential attachment rule is on; recall that the probability of attachment can be undefined if all vertices are initially disconnected).

Results includes all the tables, plots and some guiding text. Results should indicate clearly the function giving the best fit in each exercise. Methods should include any relevant methods not explained in this guide (for instance, decisions that you had to made and might have an influence on the results), etc. **Please, provide tables that summarize the results of model selection for each variant of the model, indicating the cases where the best model does not match the theoretical prediction.** The discussion should include a summary of the results and your interpretation. For any theoretical results that was not confirmed by our combination of simulation and statistical analysis, you should reason or speculate why this is so. The discussion section should also include some conclusions. You may need to refer to summary tables to draw conclusions. The report should include clear answers to the exercises. Please, refer to lab session 2 for relevant issues that need to be taken into account when reporting.

To deliver: You must deliver the report explained above. The formats accepted for the report are, in principle, pdf, Word, OpenOffice, and Postscript. You also have to hand in the source code in `R` (or other languages) that you have used

[1] Engineering Statistics Handbook. 6.3.3.1. Counts Control Charts. http://www.itl.nist.gov/div898/handbook/pmc/section3/pmc331.htm

(including some minimal comments that can help the reader) and **the output files indicated in Section 1.1**.

Procedure: Submit your work through the raco platform as a single zipped file.

Deadline: Work must be delivered within 2 weeks from the lab session you attend. Late deliveries risk being penalized or not accepted at all. If you anticipate problems with the deadline, please tell us as soon as possible.

References

[Barabási et al., 1999] Barabási, A.-L., Albert, R., and Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1-2):173–187.

[Caldarelli, 2007] Caldarelli, G. (2007). *Scale-free networks: complex webs in nature and technology*. Oxford University Press.

[Ritz and Streibig, 2008] Ritz, C. and Streibig, J. C. (2008). *Nonlinear regression with R*. Springer Science & Business Media.

Appendix

k follows a right-truncated exponential distribution (with support set defined by $k = 1, 2, ..., k_{max}$) if the probability of k is

$$p(k) = ae^{-ck}, \tag{8}$$

where it is assumed that $c > 0$ for convenience. c and k_{max} are the only free parameters. The condition

$$\sum_{k=1}^{k_{max}} p(k) = 1 \tag{9}$$

gives

$$a = \frac{1}{\sum_{k=1}^{k_{max}} e^{-ck}}. \tag{10}$$

Noticing that the $\sum_{k=1}^{k_{max}} e^{-ck}$ is the sum of a geometric series and applying $c \neq 0$, it is obtained

$$a = \frac{1 - e^{-c}}{e^{-c}(1 - e^{-ck_{max}})} \tag{11}$$

and finally

$$p(k) = \frac{1 - e^{-c}}{e^{-c}(1 - e^{-ck_{max}})} e^{-ck}. \tag{12}$$

If the right-truncation is removed, $k_{max} \to \infty$ gives

$$p(k) = \frac{1 - e^{c}}{e^{-c}} e^{-ck}. \tag{13}$$

With the change of variable $\pi = 1 - e^{-c}$ the probability mass function becomes

$$p(k) = \frac{\pi}{1 - \pi}(1 - \pi)^{k}, \tag{14}$$

which corresponds to the displaced geometric distribution of lab session 2. Notice that $\pi = 1 - e^{-c}$ is a number between 0 and 1 as $c > 0$.