How the scale-free property emerges: Preferential Attachment

Prof. Sotiris Nikoletseas

University of Patras

ΥΔΑ ΜΔΕ, Patras
2020 - 2021

The emergence of the scale-free property

- The most striking difference between a random and a scale-free network is the existence of Hubs.
- On the WWW, hubs are websites with an exceptional number of links, like google.com or facebook.com; in the metabolic network they are molecules like ATP or ADP, energy carriers involved in an exceptional number of chemical reactions.
- Why do **so different** systems as the WWW or the cell converge to a similar scale-free architecture?
- To understand why, we need to first understand the mechanism responsible for the emergence of the scale-free property.

Growth and preferential attachment (I)

There are two hidden assumptions of the Erdős-Renyi model, that are violated in real networks and differentiate random from real networks:

- The random network model assumes that we have a fixed number of nodes, N. Yet, in real networks the number of nodes continually grows.
  $e.g.$ the WWW network is continually expanded by the creation of new sites.
- The random network model assumes that we randomly choose the interaction partners of a node. Yet, in most real networks new nodes prefer to link to the more connected nodes.
  $e.g.$ We all heard about Google and Facebook, but we rarely encounter the billions of less-prominent nodes that populate the Web.

There are many other differences between real and random networks, but these two play a particularly important role in shaping a network's degree distribution.

Growth and preferential attachment (II)

In summary, the random network model differs from real networks in two important characteristics:

- **Growth:** Real networks are the result of a growth process that continuously increases N. In contrast, the random network model assumes that the number of nodes, N, is fixed.
- **Preferential Attachment:** In real networks new nodes tend to link to the more connected nodes. In contrast, nodes in random networks randomly choose their interaction partners.

The Barabási-Albert Model (I)

The recognition that growth and preferential attachment coexist in real networks has inspired a minimal model called the Barabási-Albert model, which can generate scale-free networks. The process for creating such a network is as follows:

- We start with m_0 nodes and arbitrary number of edges, with the constraint that each node should have at least one link.
- At each step, a new node is connected to the network with m links to older nodes. **(Growth property)**
- The probability $\Pi\left(k_i\right)$ that the new node will connect to a node i, depends on the degree of i as follows **(Preferential Attachment property)**:

$$\Pi\left(k_i\right) = \frac{k_i}{\sum_j k_j}$$

The Barabási-Albert Model (II)

- After t timesteps, a network with $N = t + m_0$ nodes and $m_0 + mt$ links is created.
- The network has a power-law degree distribution with degree exponent $\gamma = 3$.

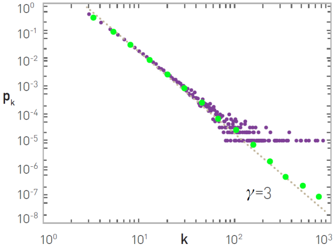

Figure: p_k *for a network of size* $N = 100.000$ *and* $m = 3$. *The linearly- binned (purple) and the log-binned version (green). The straight line has slope* $\gamma = 3$, *corresponding to the network's predicted degree exponent.*

- Hubs are created as a result of a rich-gets-richer phenomenon.

The Barabási-Albert Model - Example

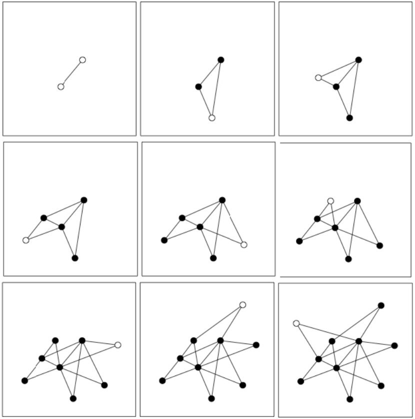

Figure: *An example of the process of creating a network using the Barabási-Albert model with $m_0 = 2$ initial nodes and $m = 2$ links added at each step.*

Degree Dynamics (I)

The time evolution of the BA model needs to be studied, in order to understand why the scale-free property appears.

- Since i has m chances to be selected every time a new node is added, the rate at which a node gets new links by new nodes connecting to it is:

$$\frac{dk_i}{dt} = m\Pi\left(k_i\right) = m\frac{k_i}{\sum_{j=1}^{N-1} k_j},$$

- the sum in the denominator obviously doesn't include the newly added node:

$$\sum_{j=1}^{N-1} k_j = 2mt - m$$

- so it is:

$$\frac{dk_i}{dt} = \frac{k_i}{2t-1}$$

Degree Dynamics (II)

- For large t the (-1) factor can be neglected:

$$\frac{dk_i}{k_i} = \frac{1}{2}\frac{dt}{t}$$

- By integrating and taking into account that $k_i\left(t_i\right) = m$, we get:

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\beta},$$

  β is called the dynamical exponent and has the value $\beta = \frac{1}{2}$ and t_i is the timestep at which node i was added to the network.

- All nodes follow the same dynamical law for increasing their degree with the same dynamical exponent $\beta = \frac{1}{2}$.

Degree Dynamics(III)

- The fact that each new node has more nodes to connect to than the previous nodes, results to the sublinear growth of the nodes' degree ($\beta < 1$).
- The earlier node i was added, the higher its degree $k_i(t)$. That's why hubs are large, they follow a phenomenon called first-mover advantage in marketing and business.
- The rate of receiving new links for a node is:

$$\frac{dk_i(t)}{dt} = \frac{m}{2}\frac{1}{\sqrt{t_i t}}$$

- This means that:
  - older nodes get more links, since they have smaller t_i.
  - all nodes get less links as time goes by, because of the $\sqrt{t}$ in the denominator.

Degree Dynamics (IV)

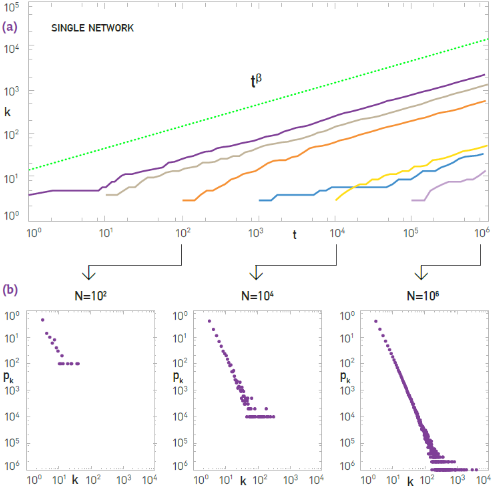

Figure: *(a)The growth of the degrees of nodes added for the different time moments, (b) Degree distribution of the network for different number of nodes added*

Degree Distribution (I)

- The power-law degree distribution of the Barabási-Albert model generated networks is their distinguishing feature.
- We can show that,

$$p(k) \approx 2m^{1/\beta}k^{-\gamma}, \text{ with } \gamma = \frac{1}{\beta} + 1 = 3$$

- This shows a connection between the degree exponent, γ, and the dynamical exponent, β, which reveals a deep relationship between the network's topology and temporal dynamics.

Degree Distribution (II)

- To get the the exact degree distribution of the Barabási-Albert model we use the following equation:

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)}$$

- γ is independent of m and m_0 parameters.
- The degree distribution is independent of both t and N. This is true for systems described by real networks, that are indeed independent of age and size.
- This, also, describes why different networks in age and size share the same degree distribution.

Degree Distribution (III)

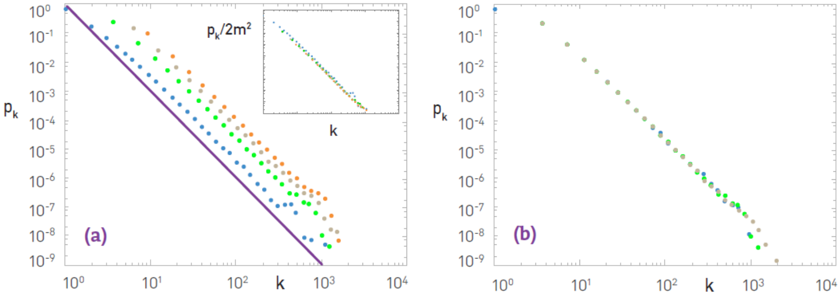

Figure: *(a) Increasing the parameters* m_0 *and* m *doesn't influence* γ*.*
$(m_0 = m = 1(blue), 3(green), 5(grey),$ *and* $7(orange))$
(b) Increasing the size of the graph doesn't change its degree distribution.
$(N = 50.000(blue), 100.000(green),$ *and* $200.000(grey),$ *with* $m_0 = m = 3)$

The absence of a property

- Are both growth and preferential attachment necessary for the scale-free property?
- To check if that's true, we will create two models each one not having one of the two properties.
- In Model A we will keep the growth property and will eliminate the preferential attachment one.
- In Model B we will keep the preferential attachment property and will eliminate the growth one.

The absence of preferential attachment - Model A

Model A keeps the growth, but eliminates preferential attachment.

- At each timestep a new node is added with m links to the already existing nodes.
- The probability that the new node is connected to a node i is independent of its degree k_i and is equal to:

$$\Pi\left(k_i\right) = \frac{1}{(m_0 + t - 1)}$$

The absence of preferential attachment - Model A

- We can show that for this Model the degree of a node i, k_i, is growing at a logarithmic rate:

$$k_i(t) = m \ln \left(e \frac{m_0 + t - 1}{m_0 + t_i - 1} \right),$$

  much slower than the power law increase.

- As a result, the degree distribution follows an exponential function:

$$p(k) = \frac{e}{m} \exp \left(-\frac{k}{m} \right)$$

- The exponential function decays much faster than the power law function, resulting to the absence of hubs.
- Consequently, the elimination of preferential attachment also eliminates the scale-free property and hence the hubs.
- In summary, the absence of preferential attachment leads to a growing network with a stationary but exponential degree distribution.

The absence of growth - Model B

Model B keeps the preferential attachment, but eliminates the growth. It starts with N nodes and then evolves as follows:

- At each timestep a random node is selected and is connected with a node i with degree k_i already in the network. The selection of i happens with probability $\Pi(k_i)$.
- Nodes' number remains constant, while the number of links increases linearly with time.
- As a result, for large t, the degree of each node increases linearly wih time:

$$k_i(t) \approx \frac{2}{N} t.$$

The absence of growth - Model B

- At an early stage, when the number of links is small, model B acts like the Barabási-Albert model with $m = 1$.
- However after a few timesteps, the degree of the nodes converges to the average degree and a peak is developed for the nodes' degree.
- p_k is not stationary ($i.e.$ It's dependent on time).
  When $t \to \frac{N(N-1)}{2}$ the graph becomes complete and all its nodes have degree $k_{max} = N - 1$, so $p_k = \delta(N - 1)$.
- In summary, absence of growth leads to the loss of stationarity, forcing the network to converge to a complete graph.

The absence of properties

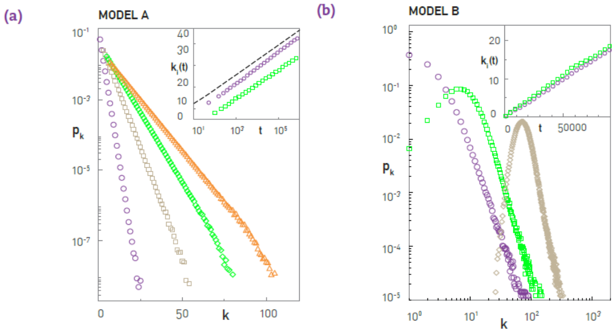

Figure: *(a) Model A: Stationary but exponential degree distribution,* $m_0 = m = 1$ *(circles), 3 (squares), 5 (diamonds), 7 (triangles) and* $N = 800.000$.
(b) Model B: Not stationary degree distribution. A peak is created after many timesteps, $N = 10.000$ *and* $t = N$ *(circles),* $t = 5N$ *(squares), and* $t = 40N$ *(diamonds).*

Measuring Preferential Attachment

- It is obvious that real networks, like WWW, have the growth property, since their size keeps getting bigger.
- On the contrary, the existence of preferential attachment is not obvious in real networks and should be detected.
- It can be measured in real networks, using the function $\Pi(k)$.
- Two hypotheses make up preferential attachment:
  1. The probability $\Pi(k)$ of connecting to a node depends on its degree k, unlike random networks.
  2. $\Pi(k)$ is linear in k.
- Both hypotheses can be tested by measuring $\Pi(k)$.

Measuring Preferential Attachment

- We get two different maps of a network at times t and $t + \Delta t$ respectively and measure $\Delta k_i = k_i(t + \Delta t) - k_i(t)$ for the nodes that changed degree during Δt.
- The relative change $\Delta k_i / \Delta t$ should follow: $\frac{\Delta k_i}{\Delta t} \sim \Pi\left(k_i\right)$.
- Since the curve of $\Delta k_i / \Delta t$ can be noisy, we measure the cumulative preferential attachment function:

$$\pi(k) = \sum_{k_i=0}^{k} \Pi\left(k_i\right)$$

- When preferential attachment is absent, we get $\Pi\left(k_i\right) =$ constant and so $\pi(k) \sim k$.
- If linear preferential attachment is present, $i.e.$ if $\Pi\left(k_i\right) = k_i$, we expect $\pi(k) \sim k^2$.

Evidence of Preferential Attachment

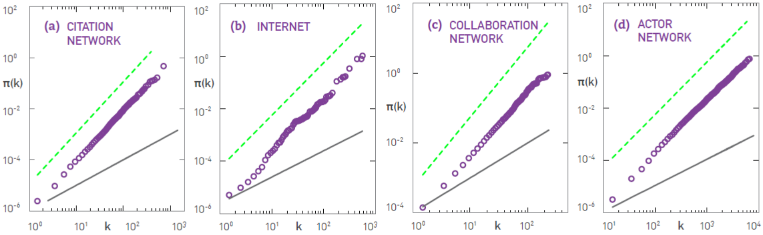

Figure: *The dashed line corresponds to linear preferential attachment ($\pi(k) \sim k^2$) and the continuous line indicates the absence of preferential attachment ($\pi(k) \sim k$).*

In line with Hypothesis 1 we detect a k-dependence in each dataset. Yet, in (c) and (d) $\pi(k)$ grows slower than k^2, indicating that for these systems preferential attachment is sublinear, violating Hypothesis 2.

Evidence of Preferential Attachment

- For each of the four systems a faster than linear increase in $\pi(k)$ is observed, indicating the presence of preferential attachment.
- $\Pi(k)$ can be approximated with

$$\Pi(k) \sim k^{\alpha}$$

.

- For the Internet and citation networks it is $\alpha \approx 1$, indicating that $\Pi(k)$ depends linearly on k. This is in line with Hypotheses 1 and 2.
- For the co-authorship and the actor network, the best fit provides $\alpha = 0.9 \pm 0.1$ indicating the presence of a sublinear preferential attachment.

Non-linear Preferential Attachment

- How does the observed sublinearity of preferential attachment affect the network's topology?
- In order to answer this, we use $\Pi(k) \sim k^{\alpha}$ instead of linear preferential attachment and observe the degree distribution of the nonlinear Barabási-Albert model.
- For $\alpha = 0$, we have no preferential attachment, getting Model A discussed earlier.
- For $\alpha = 1$, we get the Barabási-Albert model, a scale-free network with degree distribution.
- What happens for $\alpha \neq 0$ and $\alpha \neq 1$?

Sublinear Preferential Attachment $(0 < \alpha < 1)$

- For any $\alpha > 0$, more connected nodes are favored, but for $\alpha < 1$ this bias is weak, unable to cause the scale-free property.
- The degree distribution follows instead the following exponential function:

  $$p_k \sim k^{-\alpha} \exp\left(\frac{-2\mu(\alpha)}{\langle k \rangle (1-\alpha)} k^{1-\alpha}\right),$$

  where $\mu(\alpha)$ is weakly dependent on α. This means a very limited amount and size of hubs.
- This sublinearity also affects the maximum degree, k_{max}.
- The k_{max} of a scale-free network, scales polynomially with time, but for sublinear preferential attachment it is:

  $$k_{\max} \sim (\ln t)^{1/(t-\alpha)},$$

  a logarithmic growth that causes the small size of hubs.

Sublinear Preferential Attachment $(\alpha > 1)$

- For $\alpha > 1$ new nodes connect to high degree nodes more.
- For $\alpha > 2$ this behavior is obvious, where new nodes start to connect only to strong hubs.
- For $(1 < \alpha < 2)$ this behavior remains, but is not so obvious.

Evidence of Preferential Attachment

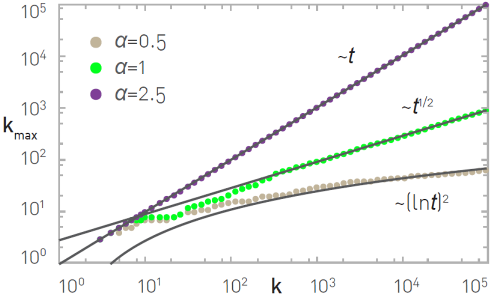

Figure: *The degree of the most connected node for different values of α.*

The origins of preferential attachment

- Where does preferential attachment come from?
- This question can be divided into two narrower questions:
  - Why does $\Pi(k)$ depend on k?
  - Why is the dependence of $\Pi(k)$ linear in k?
- There are two different groups of mechanisms that answer these questions:
  - Local mechanisms: do not require global knowledge of the network, rely on random events.
  - Optimized or global mechanisms: each new node or link balances conflicting needs, hence they are preceded by a cost-benefit analysis.

Local Mechanisms - Link Selection Model

- Although local mechanisms generate scale-free networks, without preferential attachment explicitly, they generate it implicitly.
- The link selection model is the most simple example of such a mechanism and it is defined as follows:
  - Growth: A new node is added at each timestep.
  - Link Selection: A link is selected randomly and the new node is connected to one of its two ends equiprobably.
- The model doesn't have knowledge about the topology of the network and so it is local and random.
- In contrast to the Barabási-Albert model, this model doesn't have a built-in $\Pi(k)$ function. However, it generates preferential attachment.

Link Selection Model - Generating Preferential Attachment

- The probability q_k of a node at the end of a random link having degree k ($i.e.$ the probability to find a degree$-k$ node at the end of a random link) is:

$$q_k = Ckp_k,$$

- C can be calculated using $\Sigma q_k = 1$, getting $C = 1/\langle k \rangle$, hence:

$$q_k = \frac{kp_k}{\langle k \rangle}$$

- This means that:
  - The higher the degree of a node, the more probable it is that it is located at one end of the random link.
  - The more nodes there are in the network with degree k (the higher p_K) the more likely it is that a node of degree k will be at the end of the random link.
- The fact that q_k is linear in k shows that a scale free network is built by generating linear preferential attachment.

Local Mechanisms - Copying Model (I)

- The copying model mimics a simple phenomenon: The authors of a new webpage tend to borrow links from webpages with related topics.
- A new node is added at each timestep.
- We randomly select a node u, which for example corresponds to a web document that is related to the document of the new node.
  - **Random Connection:** With probability p the new node is connected to the randomly selected node-document.
  - **Copying:** With probability $1 - p$ an outgoing link of u is selected randomly and the new node is connected to the link's target. $i.e.$ The new page copies the link of u and connects to its target.

Local Mechanisms - Copying Model (II)

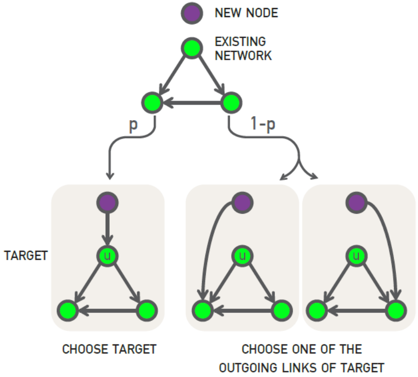

Figure: *The steps of the copying model*

Local Mechanisms - Copying Model (III)

- The probability of selecting a certain node during the Random Connection step is $1/N$, where N the number of the network's nodes.
- The probability of selecting a degree$-k$ node during the Copying step is $k/2L$ for undirected networks, where L is the number of the network's links.
- Combining these two probabilities, we get the probability of connecting to a node of degree k:

$$\Pi(k) = \frac{p}{N} + \frac{1-p}{2L}k$$

- The fact that $\Pi(k)$ is linear to k, shows that preferential attachment is generated.

Local Mechanisms - Copying Model (IV)

The copying model is popular due to its similarity to real networks like:

- **Social networks:** We “copy” the friends of our friends. It’s hard to make new friends if you don’t already have friends.
- **Citation networks:** Scientists decide what to read and cite by “copying” what other scientists have cited, since it’s impossible to read every paper available.
- **Protein Interactions:** The procedure of gene duplication, which creates new genes in cells is similar to the copying model, explaining why protein interaction networks are scale-free.

Thus, both the Link Selection and the Copying model generate linear preferential attachment.

Optimized mechanisms - Optimization (I)

- Rational choice theory in economics suggests that humans make rational decisions, balancing cost against benefits.
- Such rational decisions can lead to preferential attachment, as we will see below.
- When adding a new node to the Internet (a router), we want good bandwidth (be close to the central node) with low cost physical connection (small cable).
- These can be two conflicting goals, since the closest node (small cable) may not offer the best network performance (which is accomplished by being close to the central node).

Optimized mechanisms - Optimization (II)

- Let us assume that all nodes are located on a continent with the shape of a unit square.
- We add a new node at each timestep and place it at a random point on the plane.
- We use the following cost function to decide where to connect the new node i:

$$C_i = \min_j \left[\delta d_{ij} + h_j\right]$$

- It compares the cost of connecting to each node j already in the network.
- d_{ij} is the Euclidean distance between the new node i and the potential target j.
- h_j is the network-based distance of node j to the first node of the network. This node is considered as the “center” of the network which gives the best network performance.

Optimized mechanisms - Optimization (III)

Three distinct networks emerge for different values of δ and N:

- **Star Network** $\delta < (1/2)^{1/2}$: When $\delta = 0$, we only consider the distance from the central node, so everyone connects to it and a star network is formed. Generally, we get a star formation when h_j dominates over δd_{ij}.
- **Random Network** $\delta \geq N^{1/2}$: When δ is very large, and δd_{ij} dominates over h_j, the new nodes connect to the nodes closest to them, leading to a graph with bounded degree (no hubs) distribution like a random graph.

Optimized mechanisms - Optimization (IV)

- **Scale-free Network** $4 \leq \delta \leq N^{1/2}$: Intermediate values of δ lead to scale-free networks. The power law distribution originates from two mechanisms:
  - Optimization: each node j has a basin of attraction, whose size is correlated to h_j, which is correlated to the node's degree. If the new node falls in the basin of node j, it is connected with node j.
  - Randomness: the new node is randomly added to one of the N basins. Nodes with higher degree, have larger basins. This leads to preferential attachment.

The diversity of the mechanisms of this section suggest that linear preferential attachment is present in so many and different systems precisely because it can come from both rational choice and random actions.

Optimized mechanisms - Optimization (V)

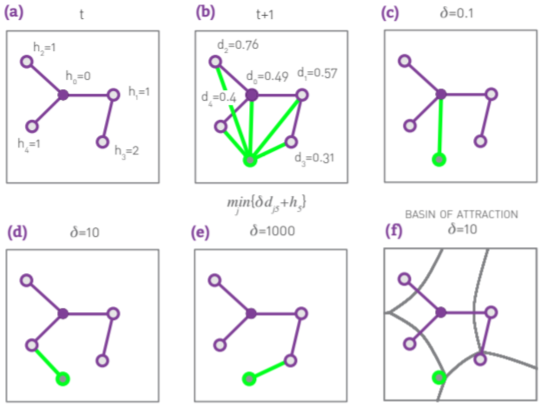

Figure: *Optimized Mechanisms*

Diameter of the Barabási-Albert model

- The network diameter for the Barabási-Albert model follows for $m > 1$ and large N:

$$\langle d \rangle \sim \frac{\ln N}{\ln \ln N}$$

- This means that the diameter grows slower than $\ln N$ and the distances are smaller than in random networks of the same size.

Diameter of the Barabási-Albert model

- The average distance has a similar behavior:

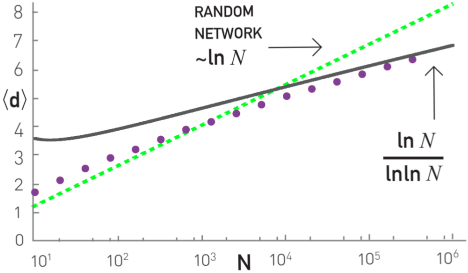

Figure: *The dependence of the average distance on the system size in the Barabási-Albert model.The continuous line corresponds to the exact result, while the dotted line corresponds to the prediction for a random network.*

- For small N the $\ln N$ term captures the scaling of $\langle d \rangle$ with N, but for large $N(\geq 10^4)$ the impact of the logarithmic correction $\ln \ln N$ becomes noticeable.

Clustering coefficient of the Barabási-Albert model

- The clustering coefficient of the Barabási-Albert model follows:

$$\langle C \rangle \sim \frac{(\ln N)^2}{N}$$

- This is quite different from the 1/N dependence of the random networks.
- The main difference is the $(\ln N)^2$ term, which increases the clustering coefficient for large N.
- Consequently the Barabási-Albert network is locally more clustered than a random network.

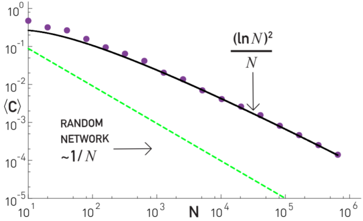

Figure: *The size N of the graph vs the average clustering coefficient for the BA and random network model*

Final Note

- The most important message of the Barabási-Albert model is that network structure and evolution are inseparable.
- Its aim is to capture the processes that assemble a network in the first place.
- While it is occasionally used as a model of the Internet or the cell, in reality it is not designed to capture the details of any particular real network.
- It is a minimal, proof of principle model whose main purpose is to capture the basic mechanisms responsible for the emergence of the scale-free property.

Acknowledgement

An initial version of this lecture was nicely prepared by Manolis Kerimakis, an excellent student of the 2019-2020 class of the YDA postgraduate program.