



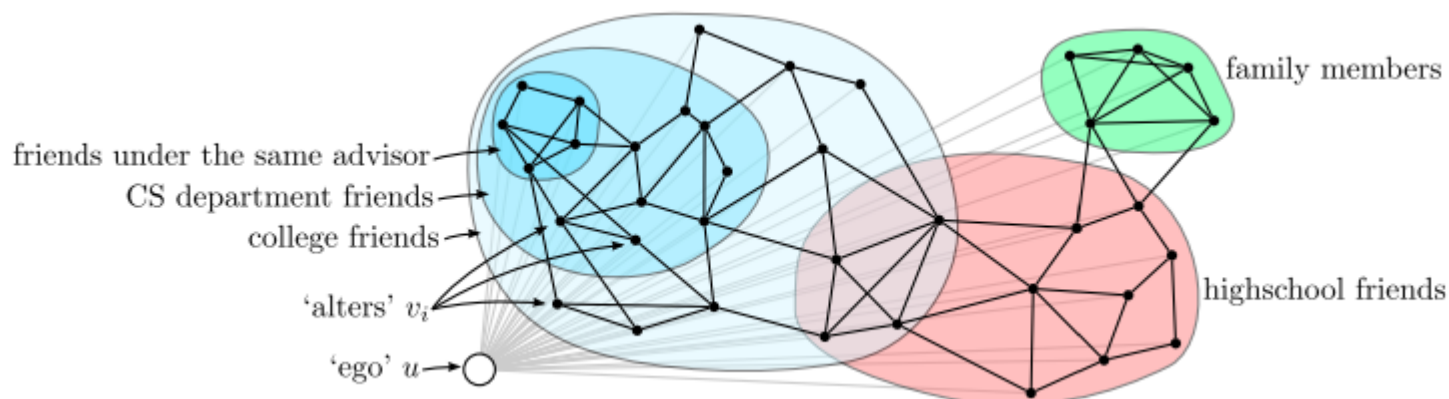
IIC-3641 Aprendizaje Automático basado en Grafos

<https://github.com/marcelomendoza/IIC3641>

- RECAPITULACIÓN -

No todas las redes tienen la propiedad scale-free (AF2)

La red que vimos en la AF2 es un subgrafo de una red social. La red que vimos era una red con un hub muy relevante (u) y el resto de los nodos siguiendo al hub. Se denominan ego-networks y son relevantes en redes jerárquicas.



Una ego-network es una red construida en torno de un usuario (u) que reúne todos sus contactos. Estas redes forman círculos anidados y pierden la propiedad scale-free.

McAuley, J., Leskovec, J. (2012)
Learning to discover social circles in
Ego Networks, NeurIPS

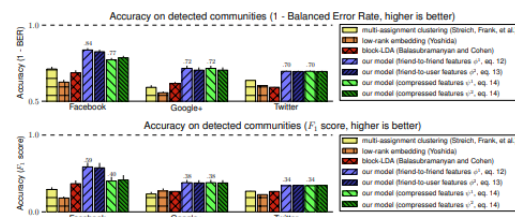


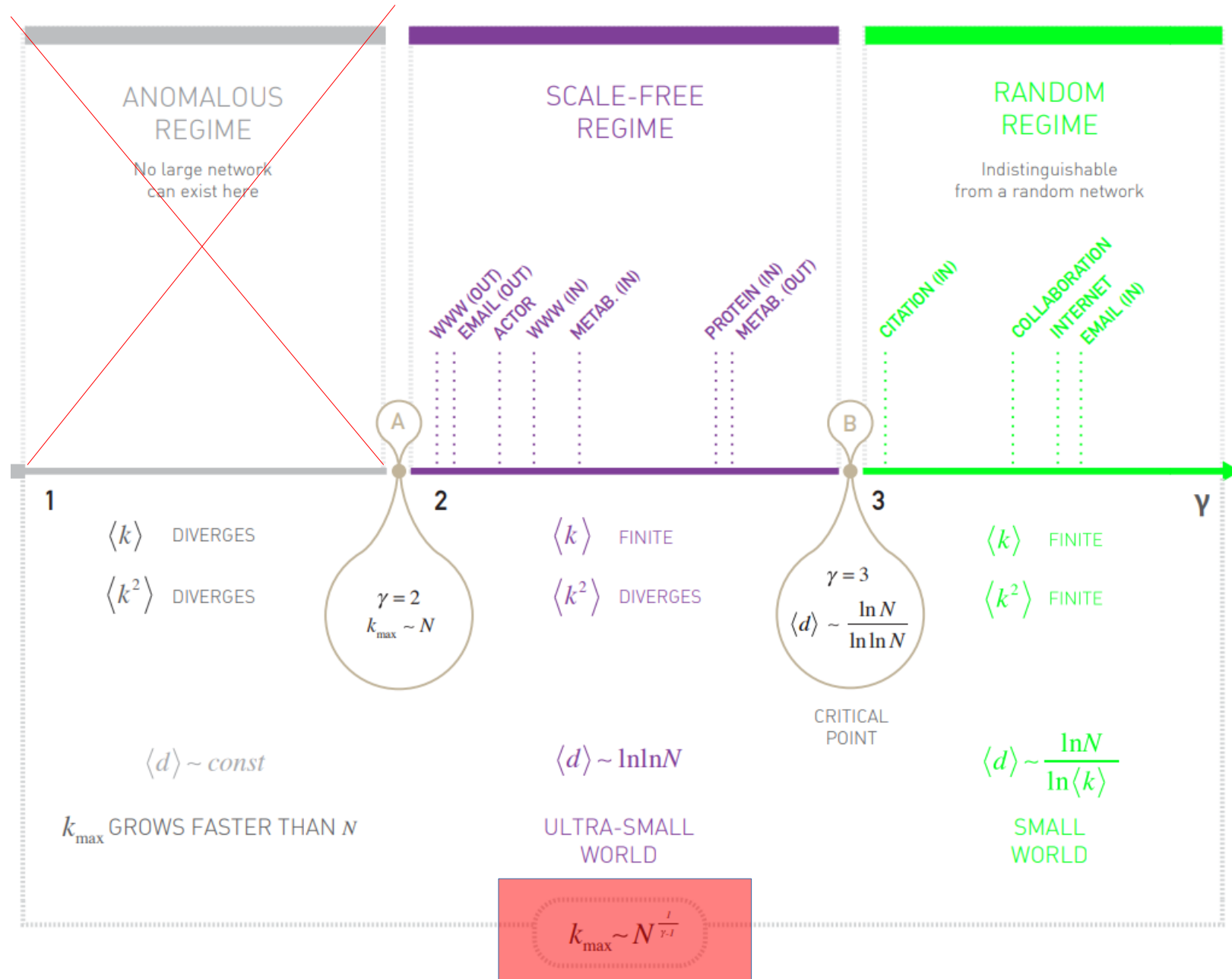
Figure 3: Performance on Facebook, Google+, and Twitter, in terms of the Balanced Error Rate (top), and the F_1 score (bottom). Higher is better. Error bars show standard error. The improvement of our best features ϕ^1 compared to the nearest competitor are significant at the 1% level or better.

Baselines. We considered a wide number of baseline methods, including those that consider only network structure, those that consider only profile information, and those that consider both. First we experimented with *Mixed Membership Stochastic Block Models* [2], which consider only network information, and variants that also consider text attributes [5, 6, 13]. For each node, mixed-membership models predict a stochastic vector encoding partial circle memberships, which we threshold to generate 'hard' assignments. We also considered *Block-LDA* [3], where we generate 'documents' by treating aspects of user profiles as words in a bag-of-words model.

Secondly, we experimented with classical clustering algorithms, such as *K-means* and *Hierarchical Clustering* [9], that form clusters based only on node profiles, but ignore the network. Conversely we considered *Link Clustering* [1] and *Clique Percolation* [21], which use network information, but ignore profiles. We also considered the *Low-Rank Embedding* approach of [30], where node attributes and edge information are projected into a feature space where classical clustering techniques can be applied. Finally we considered *Multi-Assignment Clustering* [23], which is promising in that it predicts hard assignments to multiple clusters, though it does so without using the network.

Of the eight baselines highlighted above we report the three whose overall performance was the best, namely *Block-LDA* [3] (which slightly outperformed mixed membership stochastic block models [2]), *Low-Rank Embedding* [30], and *Multi-Assignment Clustering* [23].

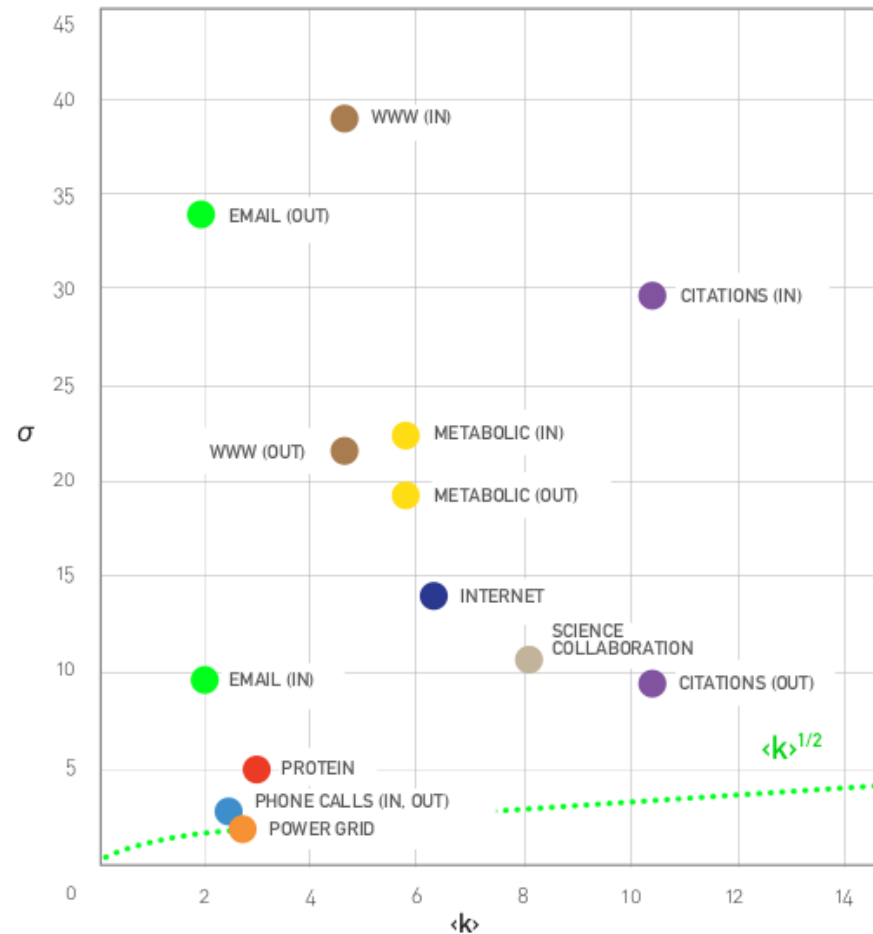
Las redes scale-free tienen grado promedio finito



Las redes scale-free tienen un grado máximo conocido

$$k_{max} = k_{min} N^{\frac{1}{\gamma-1}}$$

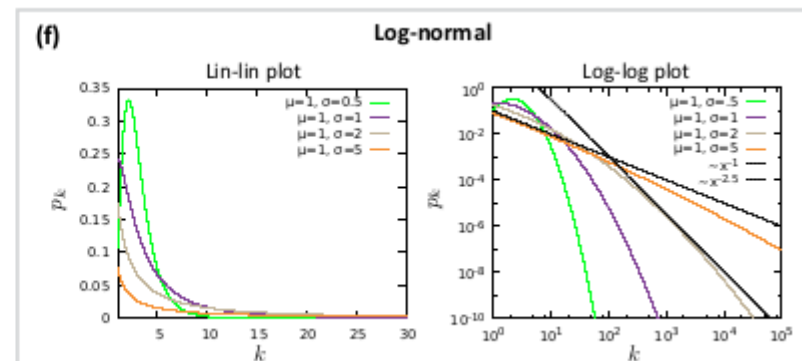
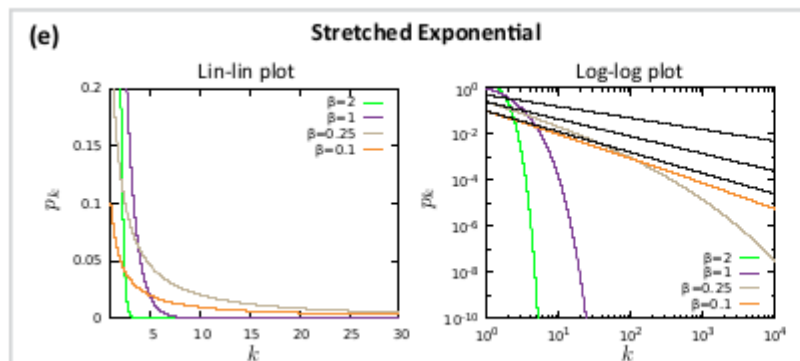
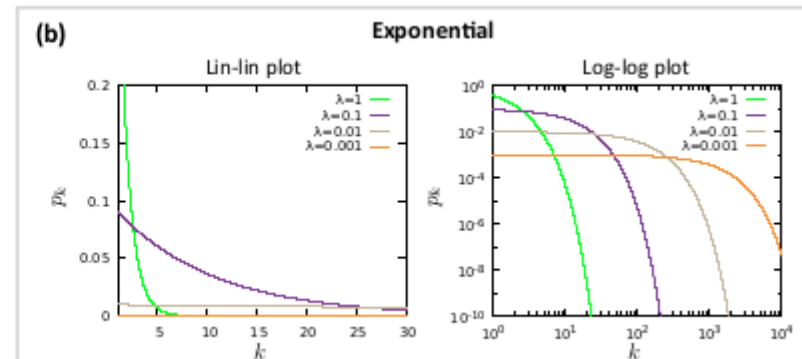
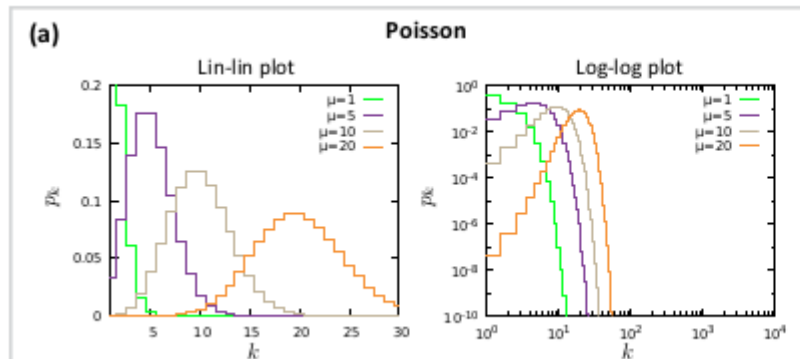
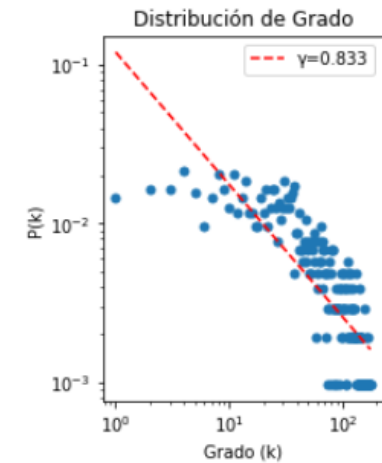
Y existen muchas redes con esta propiedad en el mundo real:



Un poco más sobre el ajuste del factor de escala

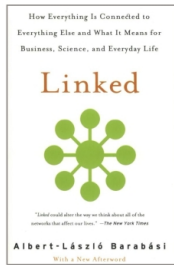
Las distribuciones de grado se ven muy distintas en log-log.
Nuestra curva de la AF2 tenía un ajuste no adecuado.

¿Se parece a alguna de estas?



- EL MODELO BARABÁSI-ALBERT -

Modelo Barabási-Albert

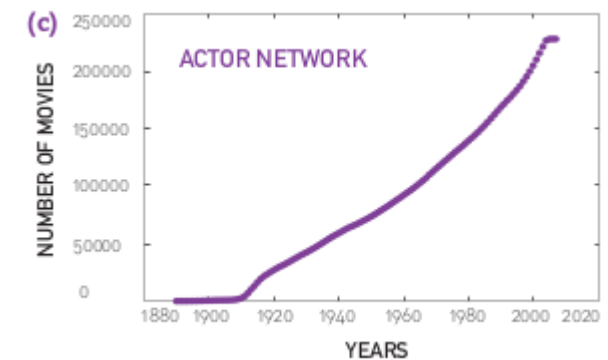
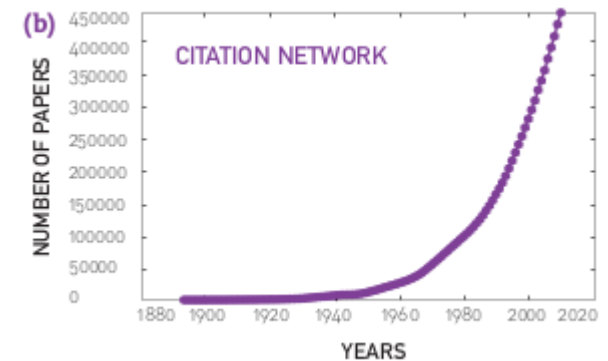
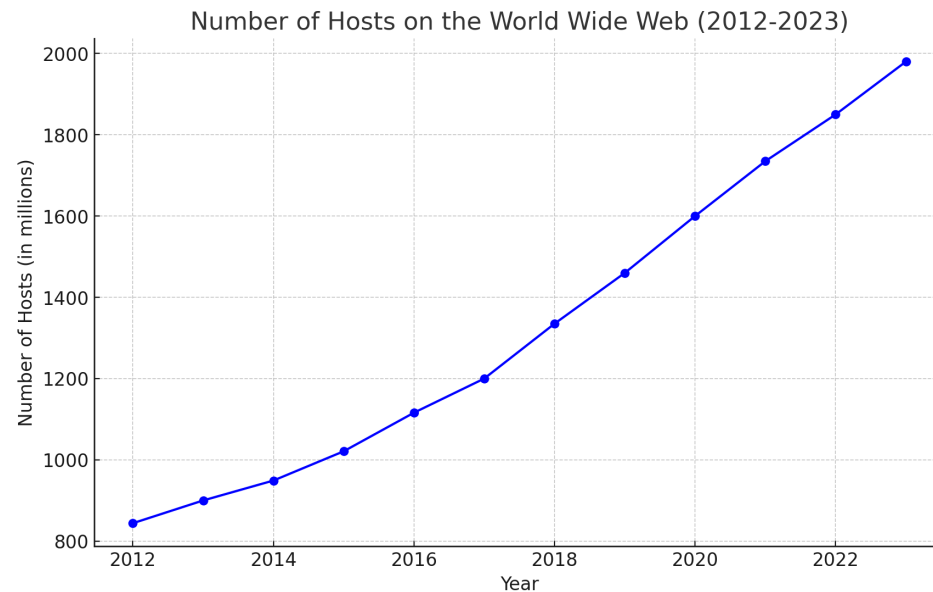


La propiedad scale-free explica el surgimiento de hubs. Muchas redes reales, como la WWW, exhiben esta propiedad.

Algo que no explica el modelo Erdős-Rényi o el Watts-Strogatz es el crecimiento de una red. Ambos modelos asumen que N está fijo.

Algunas observaciones de las redes reales:

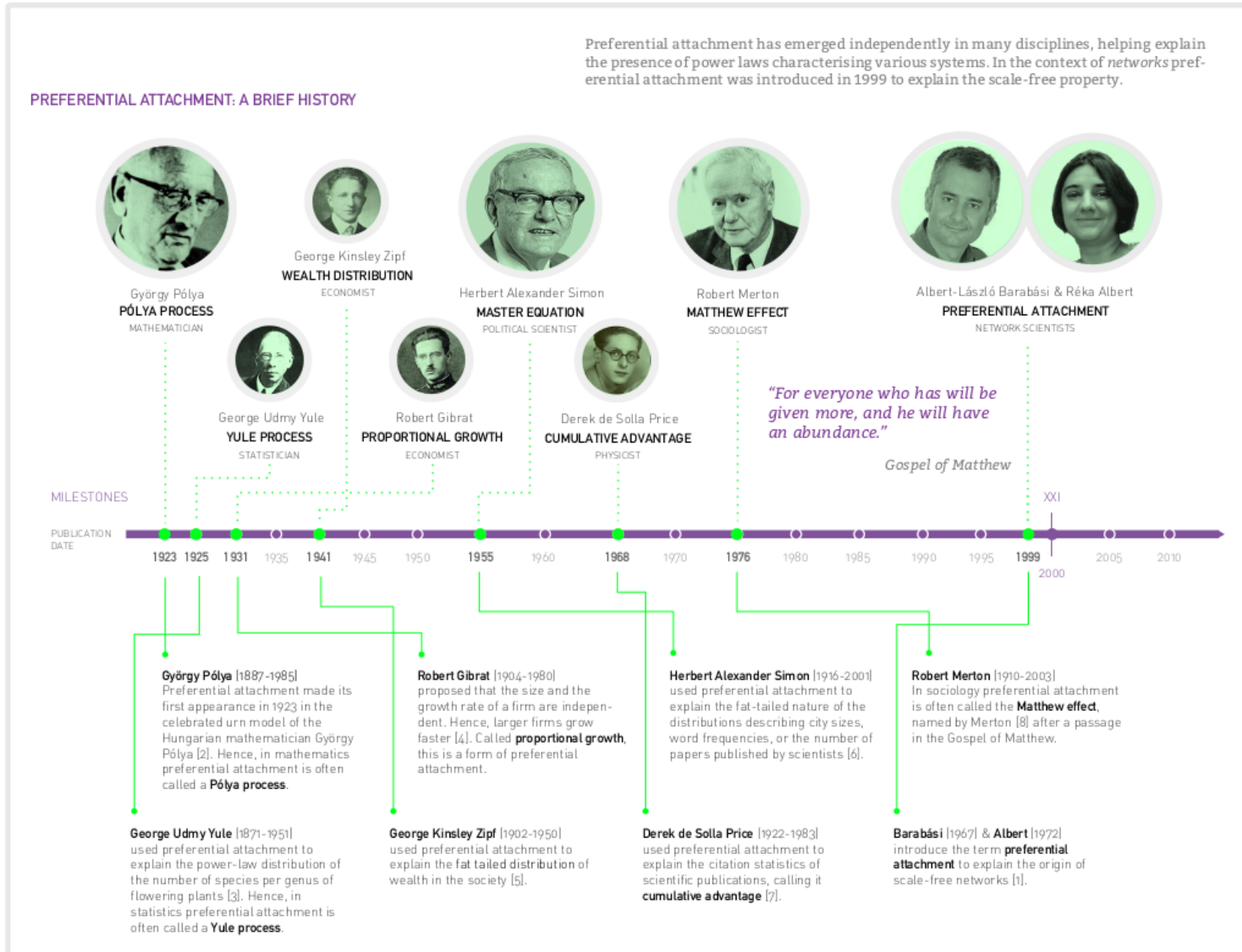
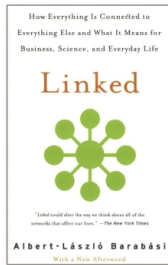
1) **Crecimiento:** Las redes crecen agregando nuevos nodos.



Modelo Barabási-Albert



2) **Preferential Attachment:** Los nodos prefieren conectarse a los nodos más conectados. Se debe al sesgo de popularidad: “Nuestro conocimiento está sesgado hacia los nodos más populares”.



Modelo Barabási-Albert

1) **Crecimiento**: En el primer step, la red tiene m_0 nodos conectados al azar. Luego, en cada timestep agregaremos un nuevo nodo a la red con $m \leq m_0$ nuevos enlaces hacia nodos preexistentes.

2) **Preferential attachment**: La probabilidad de que un enlace a un nuevo nodo se conecte al nodo i depende del grado k_i de i :

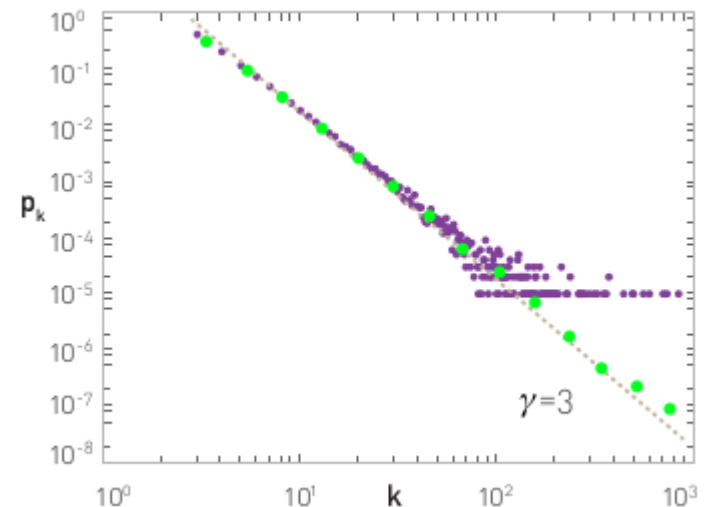
$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}$$

Notamos que después de t timesteps, el modelo Barabási-Albert produce una red con:

$$\begin{array}{ll} \# \text{ nodos:} & N = m_0 + t \\ \# \text{ enlaces:} & L = m_0 + mt \end{array}$$

Este modelo produce una distribución de grado que exhibe la propiedad scale-free.

$$N = 100,000 \text{ y } m = 3 \quad \longrightarrow$$

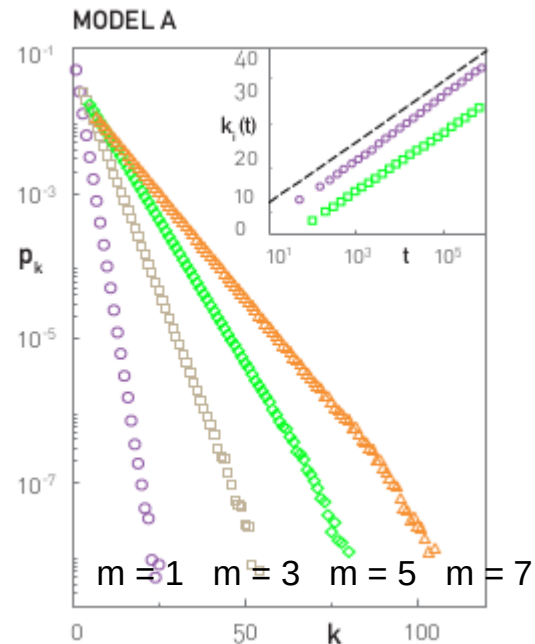


Modelo Barabási-Albert

¿Qué ocurre si desacoplamos ambos fenómenos?

Modelo A: Crecimiento sin attachment preferencial

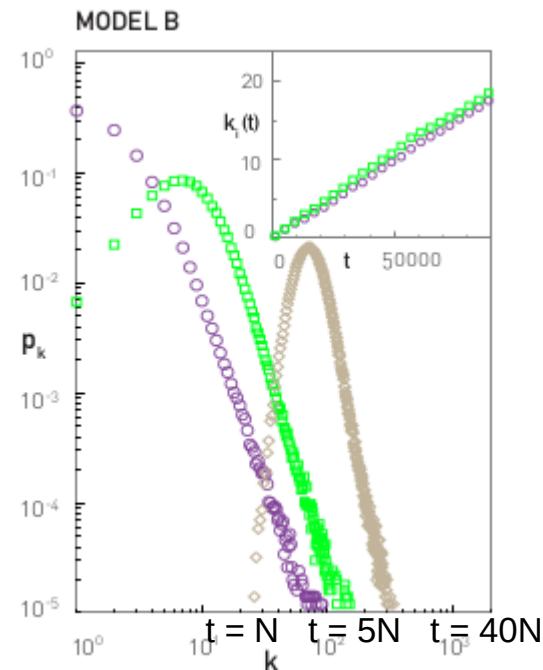
N aumenta pero los enlaces se conectan al azar con p *wiring* constante.



p_k decae más rápido que una power-law (es exponencial). Es decir, pierde la propiedad scale-free.

Modelo B: Attachment preferencial sin crecimiento.

N está fijo pero probamos para distintos timesteps



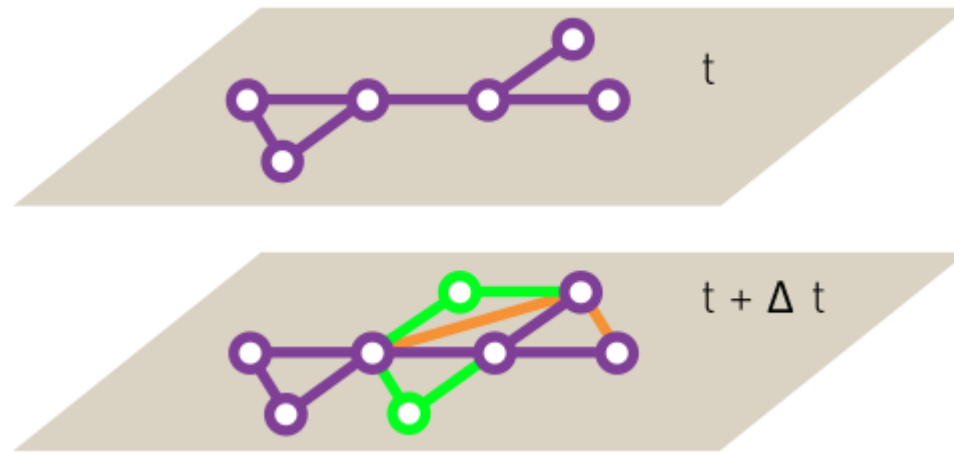
p_k cambia de forma y por tanto deja de ser estacionaria (depende de t).

Modelo Barabási-Albert

¿Cómo nos podemos asegurar que el preferential attachment existe en redes reales?

Debiera cumplirse que la forma funcional de $\Pi(k)$ sea lineal a k . Recordar que: $\Pi(k_i) = \frac{k_i}{\sum_j k_j}$

Para medir $\Pi(k)$ observaremos una red en dos momentos distintos:



Para nodos que cambiaron su grado en el frame Δt , medimos:

$$\Delta k_i = k_i(t + \Delta t) - k_i(t)$$

Modelo Barabási-Albert

Según $\Pi(k_i) = \frac{k_i}{\sum_j k_j}$, debiera ocurrir que $\frac{\Delta k_i}{\Delta t} \sim \Pi(k_i)$, esto debido a que

a mayor grado, más nuevos enlaces atraeré (fenómeno ***rich get richer***).

Dado que la curva $\Delta k_i / \Delta t$ puede ser ruidosa, analizamos la función acumulada:

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

En la ausencia de attachment preferencial tenemos que $\Pi(k_i)$ es constante y por tanto $\pi(k) \sim k$. Sin embargo, si $\Pi(k_i) = k_i$, entonces $\pi(k) \sim k^2$.

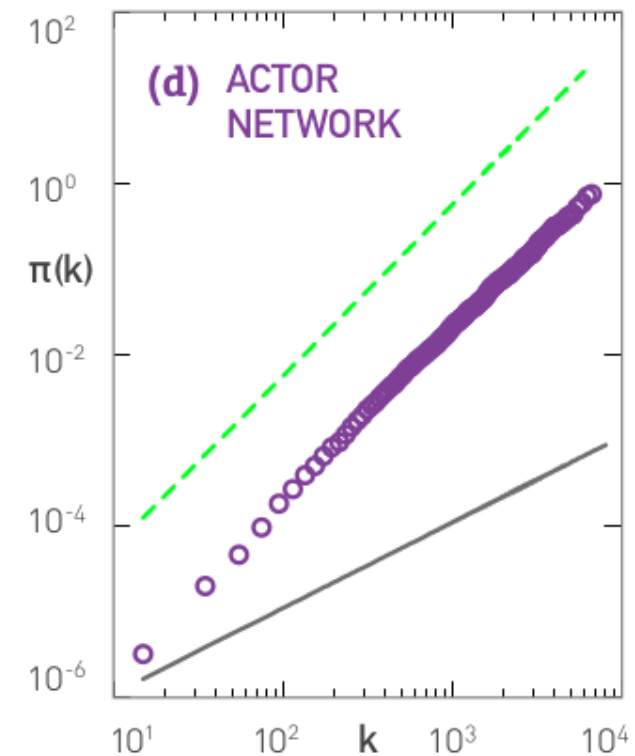
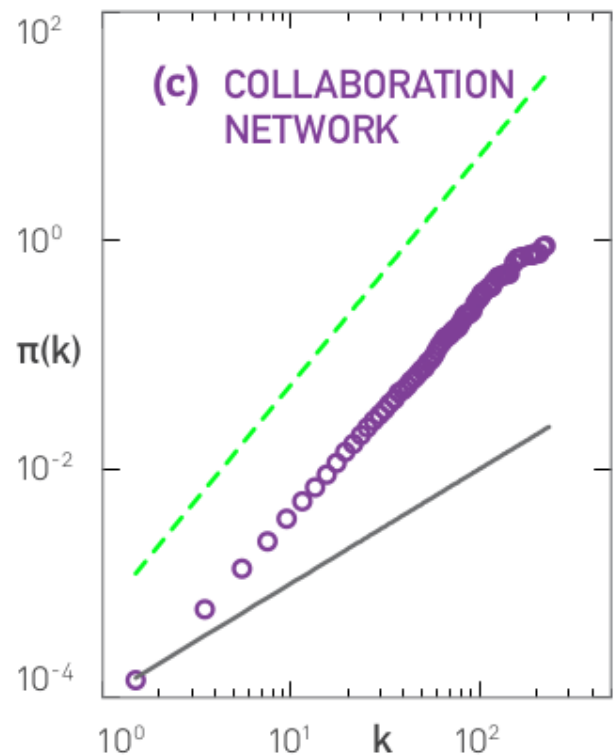
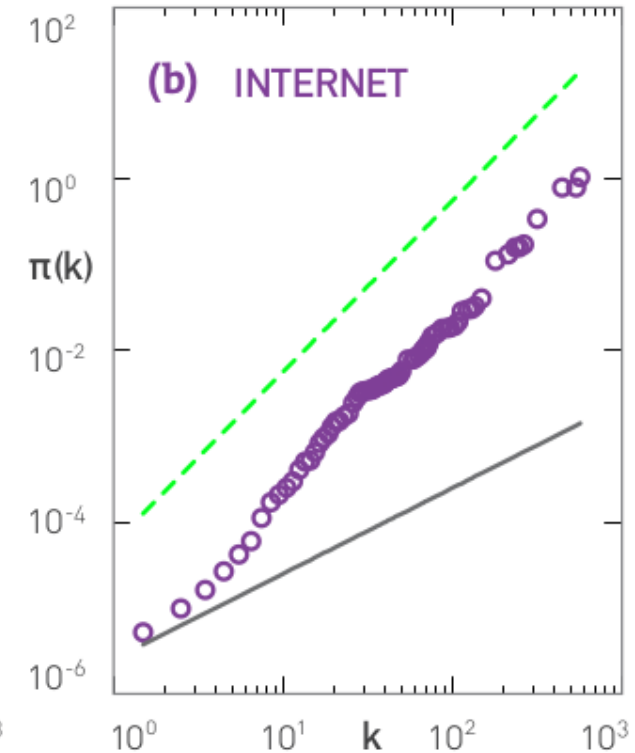
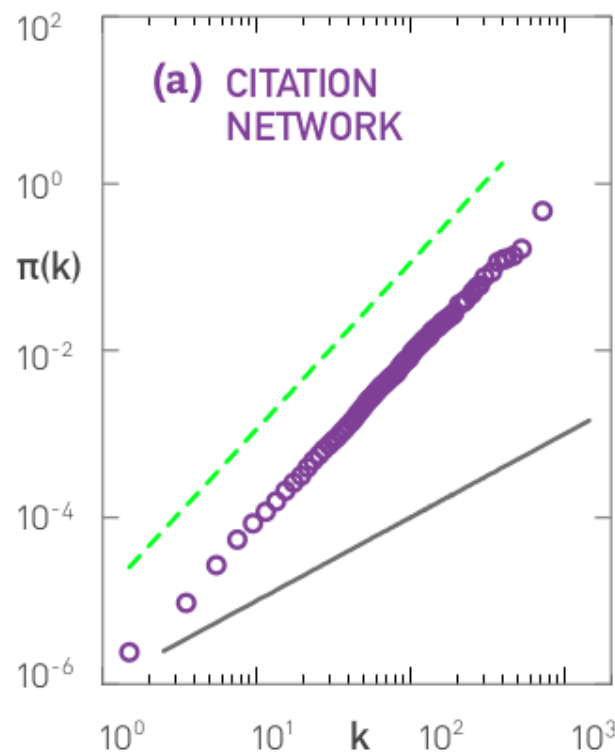
Modelo Barabási-Albert

Midamos redes reales:

Se observa que en general:

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

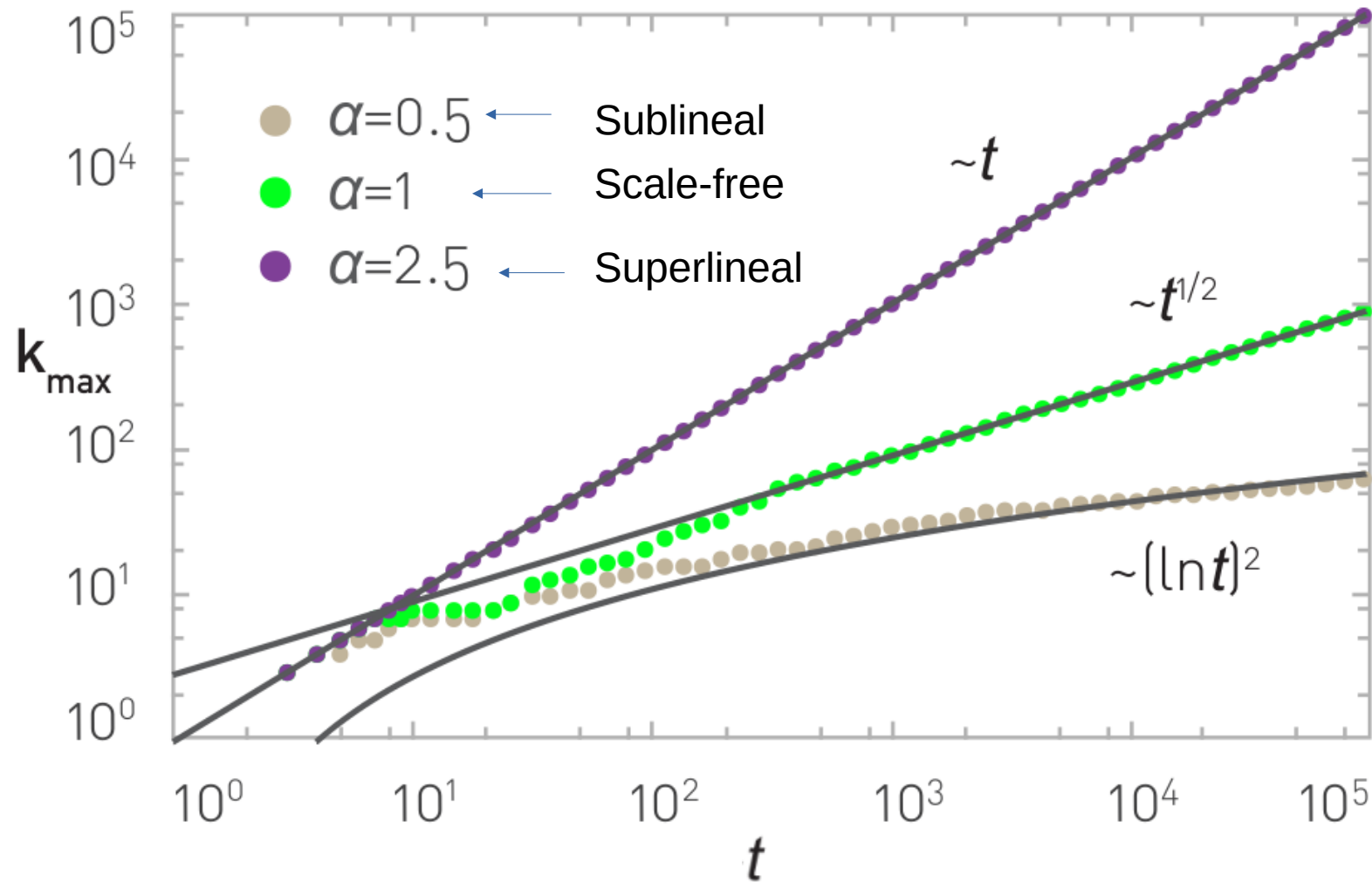
El factor alpha mide el sesgo de attachment.



Modelo Barabási-Albert

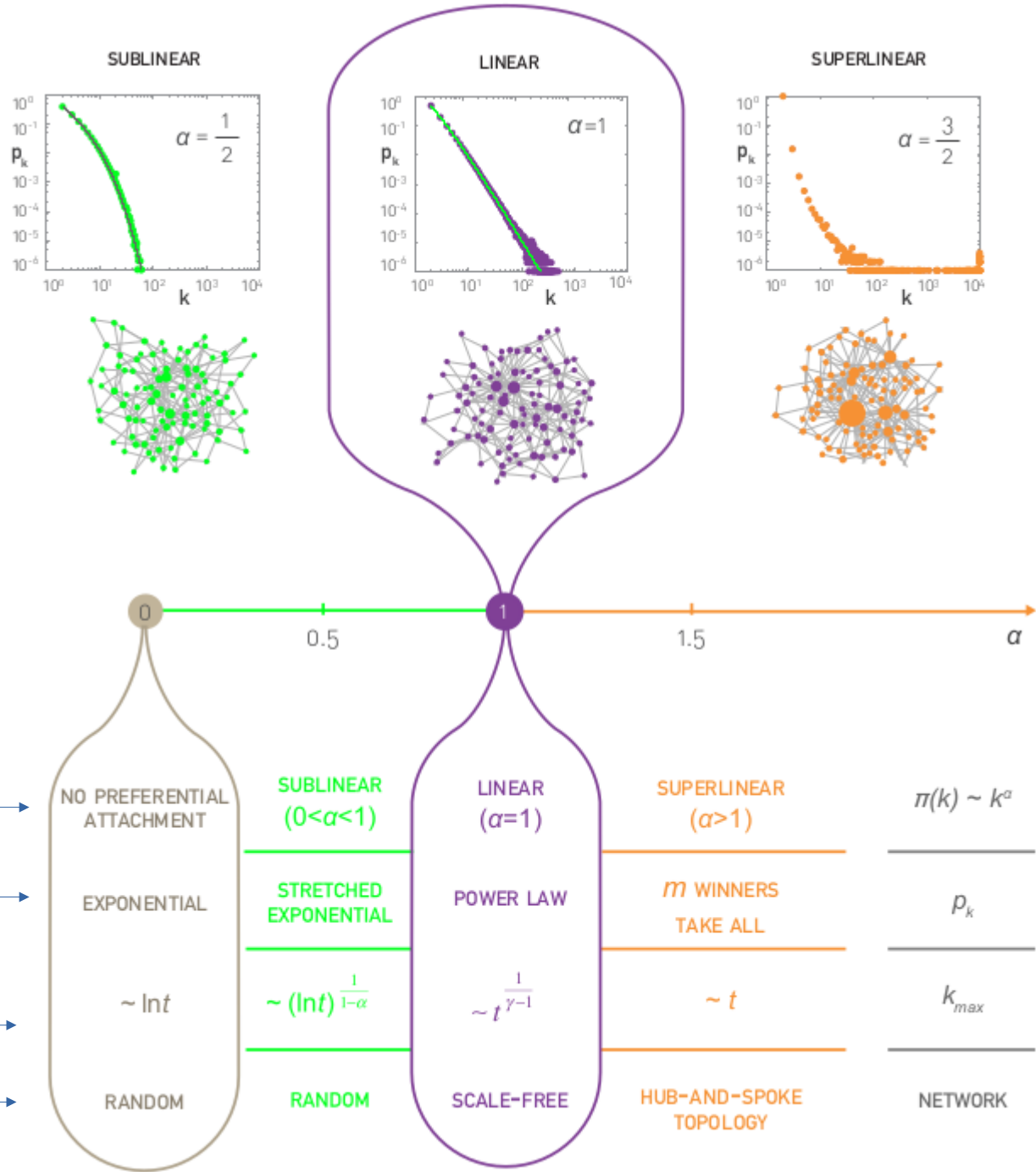
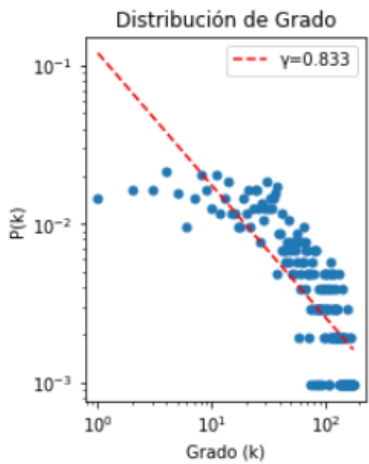
$$\Pi(k) \sim k^\alpha$$

El factor alpha determina el crecimiento de los hubs:



Modelo Barabási-Albert

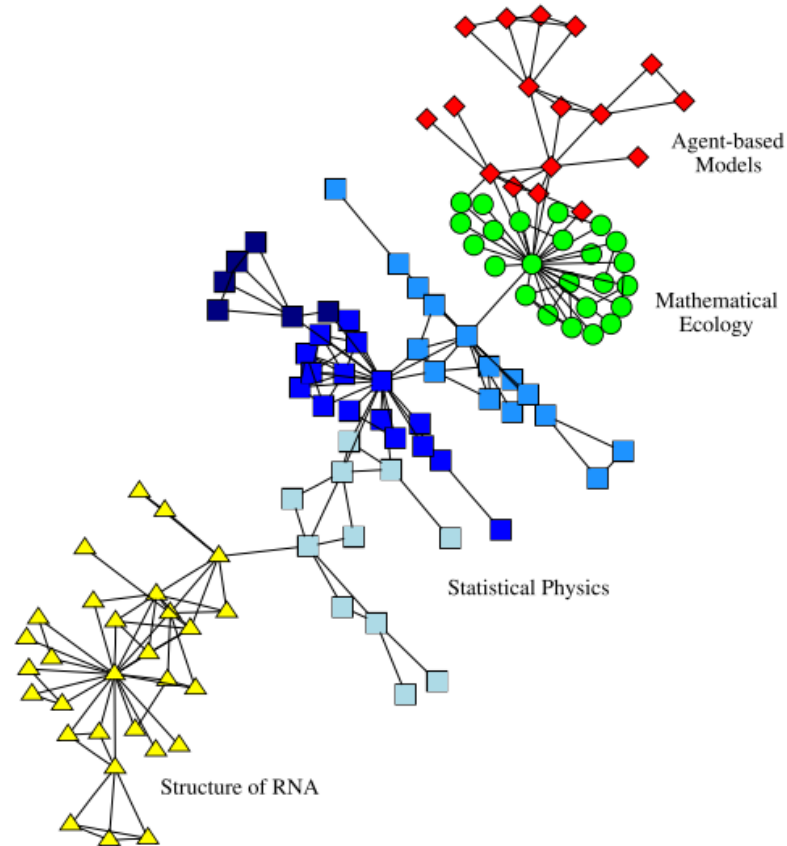
El factor alpha explica la distribución de grado:



- LA VARIANTE LANCICHINETTI – FORTUNATO – RADICCHI -

Modelo Lancichinetti – Fortunato – Radicchi (LFR)

- Las redes reales forman comunidades (se clusterizan), tienen hubs y además las comunidades tienen traslape.



Red de colaboración científica en el SFI

El modelo LFR:

- Incorpora modelado de comunidades en redes, con **heterogeneidad en la distribución de tamaños**.
- Incorpora **heterogeneidad en la distribución de grados de los nodos**.
- Ambas distribuciones siguen **ley de potencias**.

Modelo Lancichinetti – Fortunato – Radicchi (LFR)

- Modelo generativo LFR

- 1) Generar una red con nodos que sigan una ley de potencia de conectividad con exponente γ y escoger los extremos k_{\min} y k_{\max} de la distribución para obtener el grado promedio esperado $\langle k \rangle$ (empírico).
- 2) Generar tamaños de comunidades usando una ley de potencia con exponente β . Asignar los nodos a cada comunidad.
- 3) Conectar enlaces según parámetro de mezcla: una fracción $(1 - \mu)$ de enlaces de cada nodo estará con nodos de la misma comunidad, el resto hacia otros nodos. No usa preferential attachment.

Valores típicos de los parámetros de LFR:

Parámetro	Distribución	Rango típico	Valor de referencia
$\tau_1 (\gamma)$	Grados de nodos	2.0 – 3.0	2.5
$\tau_2 (\beta)$	Tamaños de comunidades	1.0 – 2.0	1.5



Benchmark graphs for testing community detection algorithms

[Andrea Lancichinetti](#), [Santo Fortunato](#), and [Filippo Radicchi](#)



Phys. Rev. E **78**, 046110 – Published 24 October, 2008

DOI: <https://doi.org/10.1103/PhysRevE.78.046110>