Gráficas aleatorias, redes sociales y el internet

Parte 2

Mariana Olvera-Cravioto

UNC Chapel Hill

molvera@email.unc.edu

30 de septiembre de 2021

Modelos de gráficas aleatorias

- Algunas redes del mundo real son demasiado grandes para ser analizadas de manera exacta.
- Algunas están cambiando constantemente.
- **Idea:** podemos conceptualizar una gráfica específica como un elemento "típico" de una familia más grande.
- Si podemos demostrar que una propiedad es cierta para una familia grande de gráficas, es probable que en particular sea cierta en nuestra gráfica.
- Las **gráficas aleatorias** son modelos matemáticos que nos ayudan a estudiar gráficas grandes del mundo real.
- Ningún modelo puede replicar todas las propiedades de una gráfica específica del mundo real, por lo que nos enfocamos en escoger modelos que comparten propiedades importantes para el problema que queremos analizar.

El límite sobre el tamaño de la gráfica

- Los modelos de gráficas aleatorias consisten en un conjunto de vértices $V_n = \{1, 2, \ldots, n\}$ y un conjunto de reglas basado en eventos aleatorios para decidir si una arista existe o no.
- Su análisis matemático se basa usualmente en tomar el **límite de gráficas crecientes**, $n \to \infty$, en la sucesión de gráficas $\{G(V_n, E_n) : n \geq 1\}$.
- Tomar el límite $n \to \infty$ simplifica los cálculos y nos permite identificar propiedades generales.
- En la práctica, establecer resultados bajo el límite $n \to \infty$ significa que las propiedades encontradas son ciertas para gráficas suficientemente grandes.

Modelos estáticos vs. modelos que evolucionan

- Los modelos de gráficas aleatorias se pueden clasificar en dos categorías: **modelos estáticos** y **modelos que evolucionan**.
- Los modelos estáticos son una "foto instantánea" de una red grande.
- En los modelos estáticos $G(V_n, E_n)$ y $G(V_{n+1}, E_{n+1})$ pueden ser completamente diferentes.
- En los modelos que evolucionan los vértices son añadidos a la gráfica (usualmente uno a la vez) para describir la manera en la que la gráfica va creciendo, por lo tanto, $G(V_n, E_n)$ y $G(V_{n+1}, E_{n+1})$ tienen muchas aristas en común.
- En muchos modelos que evolucionan, las aristas nunca desaparecen, por lo que $G(V_n, E_n)$ es una subgráfica de $G(V_{n+1}, E_{n+1})$.

El modelo Erdős-Rényi

- La gráfica aleatoria más simple es el **modelo Erdős-Rényil**.
- Consideremos una gráfica con vértices $V_n = \{1, 2, \ldots, n\}$.
- Hay $\binom{n}{2}$ posibles aristas en total, y para decidir si cada una pertenece a la gráfica vamos a echar un volado.
- Supongamos que tenemos una moneda que cae 'cara' con probabilidad $p \in (0, 1)$.
- Para cada par de vértices i y j, echamos un volado; si cae cara, dibujamos una arista entre i y j, de lo contrario no hacemos nada.
- De manera equivalente, si A denota la matriz de adyacencia de la gráfica, ponemos

$$a_{i,j} = a_{j,i} = 1(\text{el volado cae cara}), \qquad i \neq j,$$

y escribimos $a_{i,i} = 0$.

Propiedades del modelo Erdős-Rényi

- Este es el modelo de gráfica aleatorias más estudiado que hay.
- Algunas de sus propiedades son:
  - Si $np < 1$ la gráfica consiste sólo de componentes pequeños de tamaño $O(\log n)$.
  - Si $np \to c > 1$ la gráfica contiene un único componente conexo *gigante*, con todos los otros componentes de tamaño $O(\log n)$.
  - Si $np = 1$ el componente más grande es de tamaño $O(n^{2/3})$.
  - Si $p < (1-\epsilon)n^{-1}\log n$ lo más probable es que la gráfica sea **disconexa**.
  - Si $p > (1+\epsilon)n^{-1}\log n$ lo más probable es que la gráfica sea **conexa**.
- Cuando la gráfica es conexa, exhibe la propiedad del **mundo pequeño**, con distancias típicas de orden $O(\log n)$.

Distribución de los grados

- La distribución de los grados se calcula usando probabilidades binomiales.
- Fijemos un vértice $i \in V_n$, y notemos que su grado está es:

$$D_i = \sum_{j=1}^{n} \chi_{i,j}, \qquad \chi_{i,j} = 1((i,j) \in E_n)$$

- Observemos que $\chi_{i,j}$ son v.a. Bernoulli con parámetro p.
- Ya que todos los vértices tienen la misma distribución, para todo $i \in V_n$,

$$P(D_i = k) = P(D_1 = k) = P(\text{Binomial}(n,p) = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Distribución de los grados... cont.

- Más aún, si $np \to c$ cuando $n \to \infty$, podemos aproximar la binomial de la siguiente manera:

Distribución de los grados... cont.

- Más aún, si $np \to c$ cuando $n \to \infty$, podemos aproximar la binomial de la siguiente manera:

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{n^k} \cdot \frac{1}{k!}(np)^k \cdot \frac{(1-p)^n}{(1-p)^k}$$
$$\to 1 \cdot \frac{1}{k!} c^k \cdot \lim_{n\to\infty} e^{n\log(1-p)}, \qquad n \to \infty$$

- Para calcular este último límite, notemos que

$$\lim_{n\to\infty} n\log(1-p) = \lim_{n\to\infty} (-np + O(np^2)) = -c$$

- Por lo tanto,

$$\lim_{n\to\infty} P(D_1 = k) = \frac{e^{-c}c^k}{k!}, \qquad k \geq 0,$$

que es una distribución Poisson con media c.... no es **libre de escala**.

Poisson vs. libre de escala

- La distribución Poisson tiene **cola ligera**, i.e., decae exponencialmente.
- Las variables aleatorias Poisson suelen tomar valores cercanos a su media.
- Las distribuciones libres de escala tienen **cola pesada**, i.e.,

$$\sum_{k=0}^{\infty} e^{\epsilon k} P(D = k) = \infty$$

para toda $\epsilon > 0$.

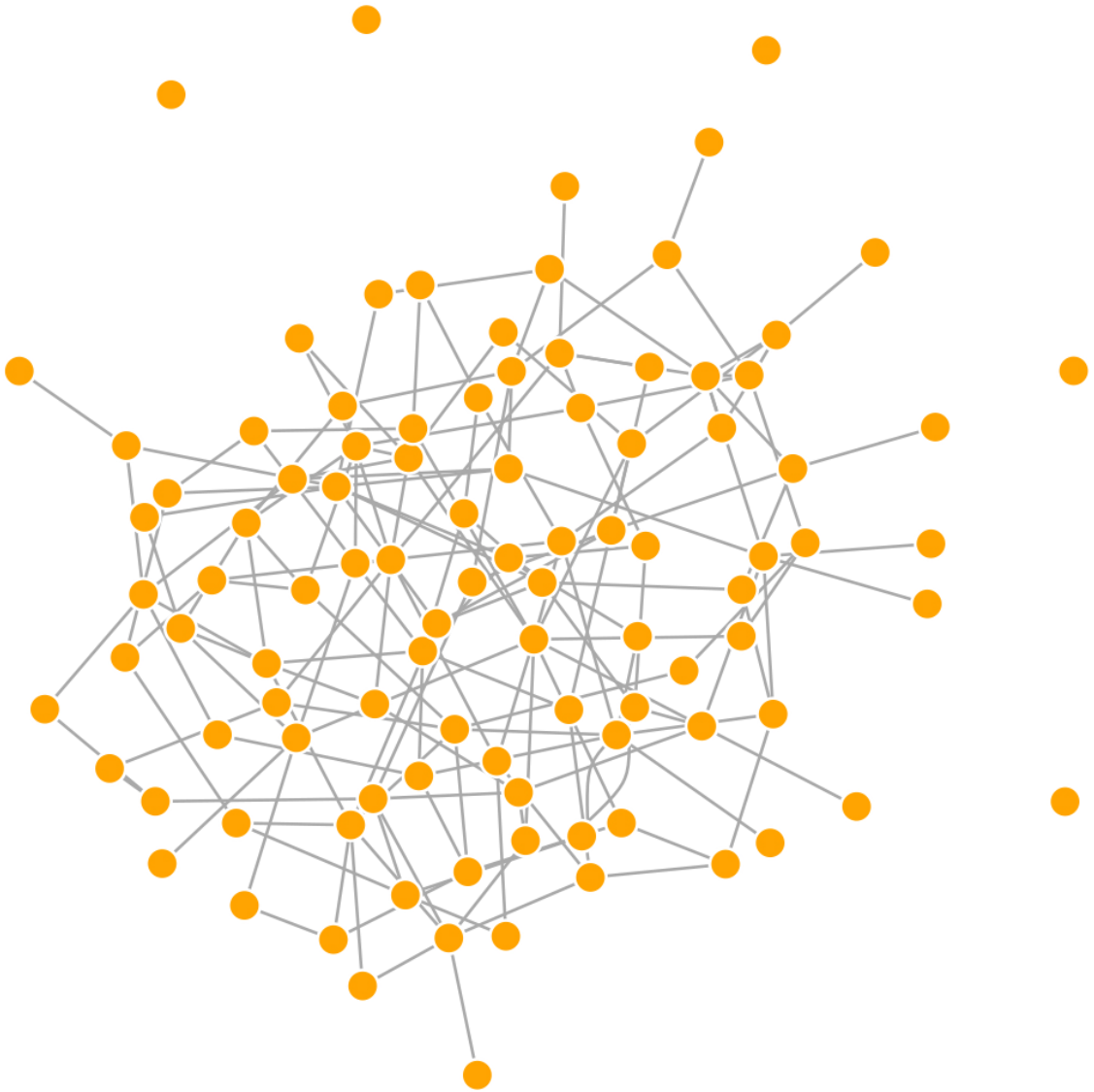
- Las variables aleatorias con colas pesadas pueden tomar valores extremadamente grandes.
- En particular, para toda $k \geq 1$,

$$\lim_{m \to \infty} P(D > k + m | D > m) = 1$$

que puede ser interpretado como:

"Dado que D es grande, lo más probable es que sea enorme."

Una gráfica Erdős-Rényi



Gráficas aleatorias inhomogéneas

- Las gráficas Erdős-Rényi son bastante **homogéneas**, i.e., todos sus vértices tienen grados cercanos a la media.
- Las redes del mundo real tienden a ser libres de escala.
- Podemos generar gráficas aleatorias con grados inhomogéneos si permitimos que las probabilidades entre aristas varien de vértice a vértice.
- A cada vértice $i \in V_n$ le asignamos un valor $w_i \geq 0$, y definimos su probabilidad entre aristas de acuerdo a

$$p_{i,j}^{(n)} := P((i,j) \in E_n) = \frac{w_i w_j}{l_n} \wedge 1, \qquad i \neq j,$$

  donde $l_n = w_1 + \cdots + w_n$.
- La matriz de adyacencia de la gráfica es:

$$a_{i,j} = \begin{cases} 1, & \text{con probabilidad } p_{i,j}^{(n)}, \\ 0 & \text{con probabilidad } 1 - p_{i,j}^{(n)}. \end{cases}$$

Gráficas aleatorias inhomogéneas... cont.

- Cada arista es independiente de todas las otras aristas.
- Esta probabilidad entre aristas se conoce como el **modelo Chung-Lu**.
- El valor esperado del grado del vértice $i \in V_n$ es:

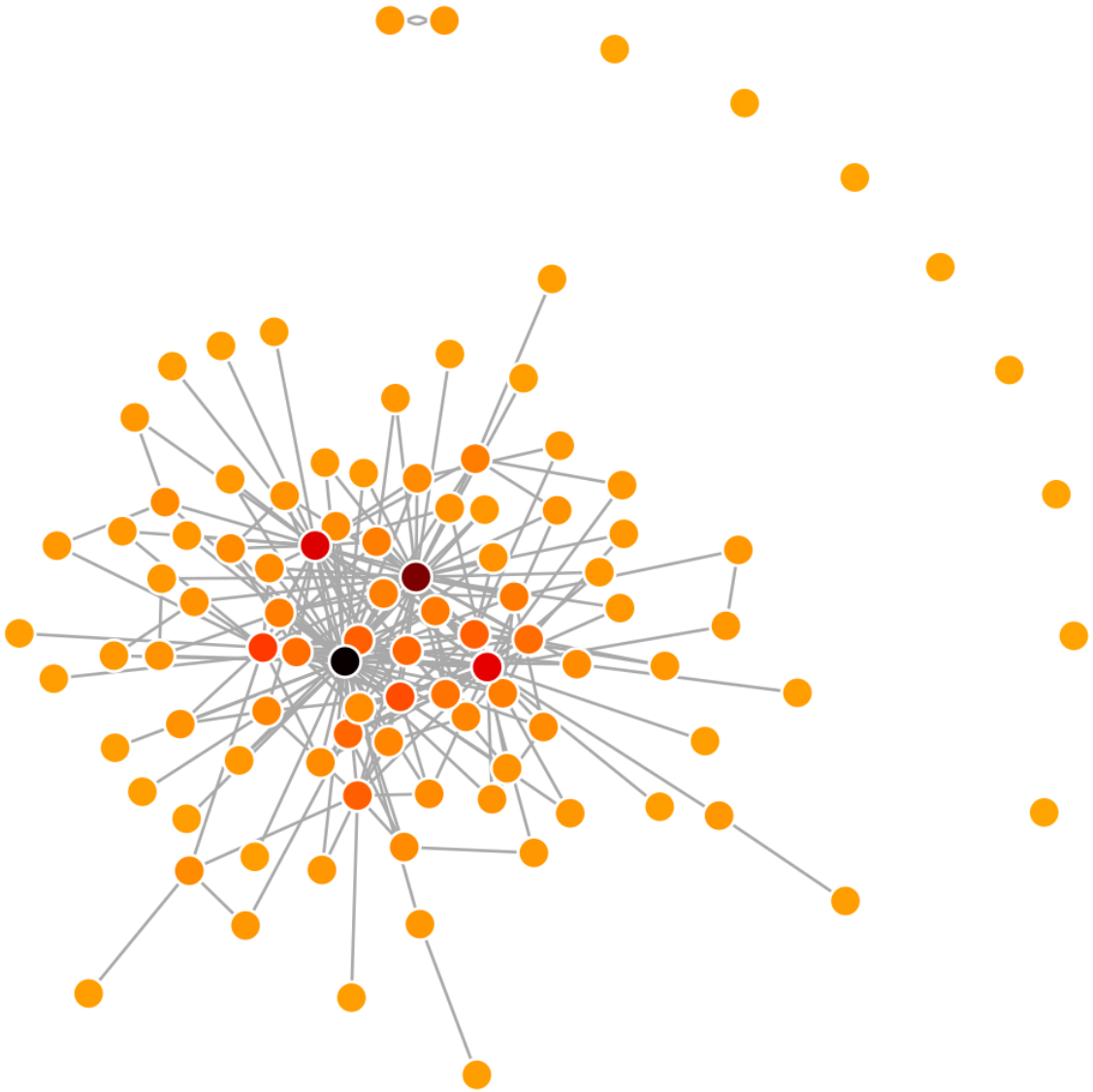
$$E[D_i] = \sum_{j=1}^{n} p_{i,j}^{(n)} \approx w_i$$

- Si definimos

$$F(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} 1(w_i \leq x),$$

  tenemos que la distribución de los grados se "parece" a F.
- Si escogemos $w_i = p$ para todo $i \in V_n$ obtenemos el modelo Erdős-Rényi.
- Para generar una gráfica libre de escala escogemos F de acuerdo a una distribución Pareto.

Una gráfica aleatoria inhomogénea



Gráficas con comunidades

- Las gráficas aleatorias inhomogéneas pueden ser **libres de escala** y van a tener la propiedad del **mundo pequeño**.
- Sin embargo, no tienen comunidades.
- Supongamos que queremos generar una gráfica con K comunidades.
- A cada vértice $i \in V_n$ le asignamos una etiqueta $c_i \in \{1, 2, \dots K\}$.
- Después sampleamos las aristas de manera independiente usando las probabilidades:

$$p_{i,j}^{(n)} = P((i,j) \in E_n) = \frac{\kappa(c_i, c_j)}{n}, \qquad i \neq j,$$

donde $\kappa : \{1, \dots, K\} \times \{1, \dots, K\} \to [0, \infty)$.

- El tamaño de la comunidad $k \in \{1, \dots, K\}$ es $n\pi_{n,k} = \sum_{i=1}^{n} 1(c_i = k)$.

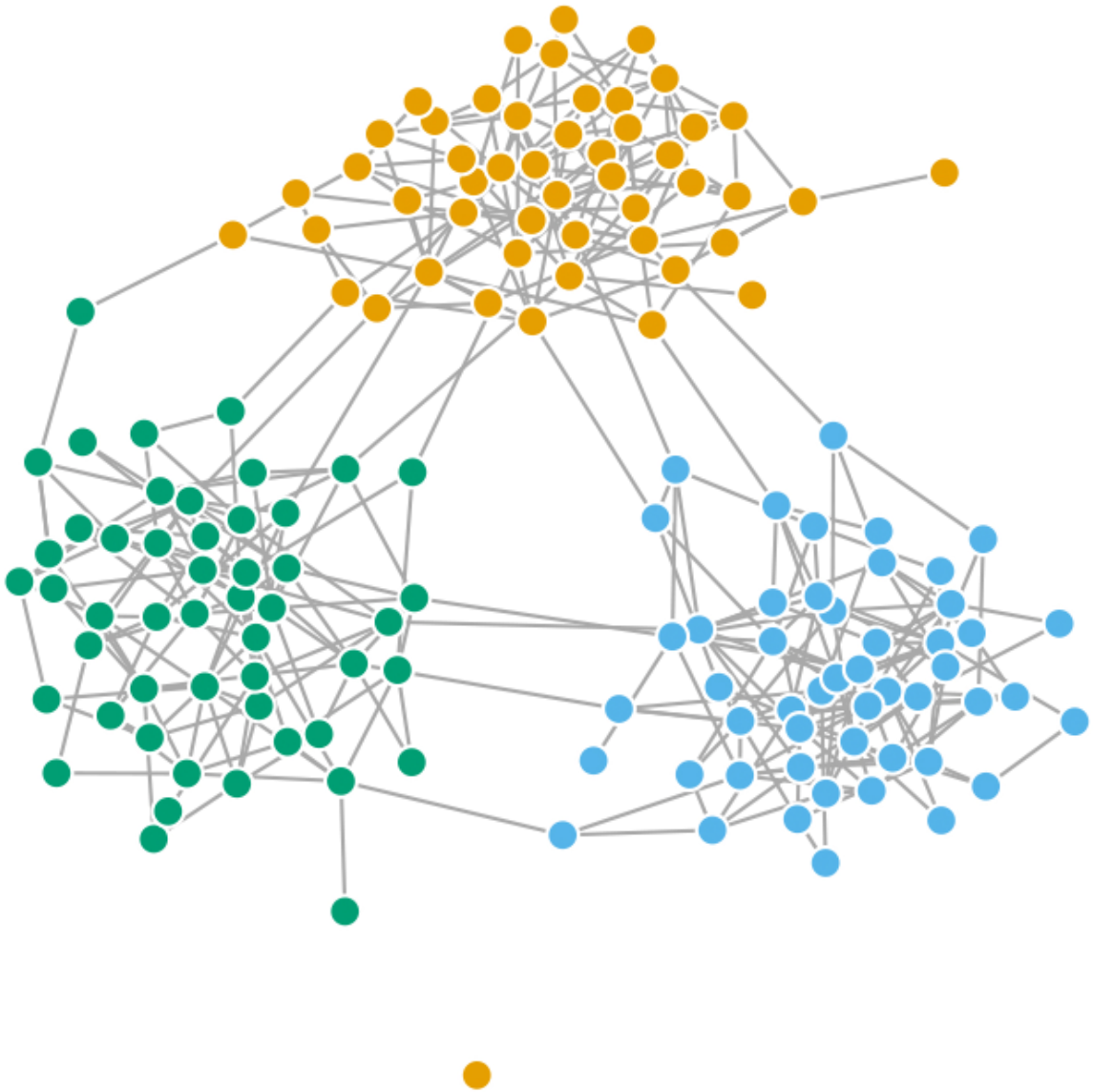
Gráficas con comunidades... cont.

- Esta construcción se conoce como el **modelo estocástico por bloques**.
- Para crear las comunidades escogemos $\kappa(c_i, c_j)$ "grande" para $i = j$, y "pequeño" pare $i \neq j$.
- El grado esperado de un vértice en la comunidad $m \in \{1, \ldots, K\}$ es:

$$E[D_i | c_i = m] = \sum_{j=1}^{n} \frac{\kappa(m, c_j)}{n} = \sum_{r=1}^{K} \kappa(m, r)\pi_{n,r}$$

- Los modelos estocásticos por bloques son homogéneos dentro de cada comunidad, pero de una comunidad a otra pueden tener grados con valores esperados diferentes.

Un modelo estocástico por bloques



Gráficas con grupos

- El **coeficiente de agrupamiento global** de una gráfica es

$$\frac{\text{número de triángulos}}{\text{número de enlaces con 3 vértices y 2 aristas}}$$

- Las gráficas aleatorias inhomogéneas no tienen agrupamiento significativo.
- De hecho, las gráficaas aleatorias inhomogéneas son **localmente árboles**.
- Tienen ciclos "largos" de longitud $O(\log n)$.
- El coeficiente de agrupamiento en los modelos que hemos visto converge a cero cuando $n \to \infty$.
- Las gráficas del mundo real tienden a tener coeficientes de agrupamiento positivos, especialmente las redes sociales.

Gráficas con grupos... cont.

- Para construir una gráfica con agrupamiento significativo, empezamos por generar una **gráfica bipartita** con conjuntos de vértices $V_n = \{1, \ldots, n\}$ y $\mathcal{A}_m = \{a_1, \ldots, a_m\}$, $n, m \geq 1$.
- A cada vértice $i \in V_n$ le asignamos un valor $w_i \geq 0$ y definimos

$$p_i = \frac{\gamma w_i}{n} \wedge 1,$$

  donde $\gamma > 0$ es un parámetro fijo.
- Después, para cada $i \in V_n$ y cada uno de los vértices en A_m, echamos un volado que cae cara con probabilidad p_i, y dibujamos una arista si el volado cae cara.
- Sea $N(i) \subseteq \mathcal{A}_m$ el conjunto de vecinos de i.
- Ahora construimos una nueva gráfica $G(V_n, E_n)$, con matriz de adyacencia A, de acuerdo a:

$$a_{i,j} = 1(N(i) \cap N(j) \neq \varnothing)$$

Gráficas con grupos... cont.

- Este modelo es conocido como una **gráfica de intersección**.
- Sea $F(x) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} 1(w_i \leq x)$ la distribución de los pesos, y supongamos que tiene media fnita.
- Si escogemos $m = \lfloor \beta n \rfloor$, el grado del vértice $i \in V_n$ en $G(V_n, E_n)$ va a tener (aproximadamente) la distribución de

$$\text{Poisson}(\beta\gamma w_i) + \text{Poisson}(\gamma),$$

  con las dos Poisson independientes una de otra.
- Como en el caso de las gráficas aleatorias inhomogéneas, podemos obterner la propiedad libre de escala si esgogemos F que sea una distribución Pareto.
- Los parámetros β, γ pueden ajustarse para obtener cualquier coeficiente de agrupamiento en el rango $(0, 1)$; valores pequeños de $\beta\gamma$ producen coeficientes más grandes.

Una gráfica de intersección

El modelo Albert-Barabási

- Todos los modelos de gráficas aleatorias que hemos visto son **estáticos**.
- Los modelos estáticos no explican como crecen las gráficas.
- Los modelos que **evolucionan** proponen un mecanismo para escoger cómo se conecta un vértice que acaba de llegar.
- Numeramos los vértices en el orden en el que se agregan a la gráfica.
- Uno de los modelos más famosos de gráficas que evolucionan es el modelo **Albert-Barabási** o de **conexión preferencial**.
- Este modelo asume que un vértice que va llegando escoge el vértice al que va a conectarse de acuerdo a una probabilidad proporcional a su grado.
- En otras palabras, los recién llegados "prefieren" conectarse a los vértices con grados grandes.

El modelo Albert-Barabási... cont.

- Este modelo comienza con un solo vértice que tiene un bucle.
- En cada paso, un nuevo vértice llega y se agrega a la gráfica ya sea conectándose con una arista a un vértice existente, o creando un bucle.
- Sea $D_i(k)$ el grado del vértice i después de que k vértices han sido agregados.
- Cuando llega el vértice $k+1$ se conecta al vértice i con probabilidad:

$$p_i(k) = \begin{cases} \frac{D_i(k)}{2k+1}, & i = 1, \ldots, k, \\ \frac{1}{2k+1}, & i = k+1. \end{cases}$$

- Este modelo produce gráficas **libres de escale** con distribución del los grados:

$$P_k(n) = \frac{1}{n} \sum_{i=1}^{n} 1(D_i(n) = k) \approx 4k^{-3}$$

cuando n es grande.

Gracias por su atención.