FACULTAD DE INGENIERÍA

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Estructura de las grandes redes de datos

Paola Bermolen

paola@fing.edu.uy

22 de agosto de 2022

1. **Componentes**
2. **Distribución de grados**
3. **Redes *Power Law* y *Scale-Free***
   - Visualizando y ajustando distribuciones *power law*
4. **Otras medidas de centralidad**

Estructura de las redes reales

- Obejtivo: aplicar los conceptos vistos hasta ahora para capturar el comportamiento de las redes reales
- ¿Qué tan conectadas son?

Estructura de las redes reales

- Obejtivo: aplicar los conceptos vistos hasta ahora para capturar el comportamiento de las redes reales
- ¿Qué tan conectadas son?
- ¿Cuál es el tamaño de la componente más grande?

Estructura de las redes reales

- Obejtivo: aplicar los conceptos vistos hasta ahora para capturar el comportamiento de las redes reales
- ¿Qué tan conectadas son?
- ¿Cuál es el tamaño de la componente más grande?
- ¿Cómo es la secuencia de grados?

Estructura de las redes reales

- Obejtivo: aplicar los conceptos vistos hasta ahora para capturar el comportamiento de las redes reales
- ¿Qué tan conectadas son?
- ¿Cuál es el tamaño de la componente más grande?
- ¿Cómo es la secuencia de grados?
- ¿Cuáles son los largos de los caminos entre sus vértices?

Estructura de las redes reales

- Obejtivo: aplicar los conceptos vistos hasta ahora para capturar el comportamiento de las redes reales
- ¿Qué tan conectadas son?
- ¿Cuál es el tamaño de la componente más grande?
- ¿Cómo es la secuencia de grados?
- ¿Cuáles son los largos de los caminos entre sus vértices?
- ¿Cuál es el diámetro de la red?

Estructura de las redes reales

- Obejtivo: aplicar los conceptos vistos hasta ahora para capturar el comportamiento de las redes reales
- ¿Qué tan conectadas son?
- ¿Cuál es el tamaño de la componente más grande?
- ¿Cómo es la secuencia de grados?
- ¿Cuáles son los largos de los caminos entre sus vértices?
- ¿Cuál es el diámetro de la red?
- ? Cuál es el coeficiente de clustering típico?

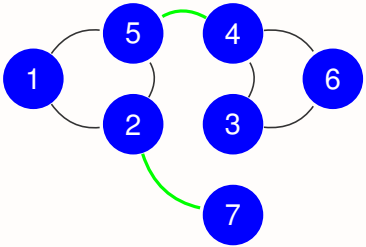
Estructura de las redes reales

	Network	Type	n	m	c	S	ℓ	α	C	C_{WS}	r	Ref(s).
Social	Film actors	Undirected	449 913	25 516 482	113.43	0.980	3.48	2.3	0.20	0.78	0.208	20, 466
	Company directors	Undirected	7 673	55 392	14.44	0.876	4.60	–	0.59	0.88	0.276	131, 369
	Math coauthorship	Undirected	253 339	496 489	3.92	0.822	7.57	–	0.15	0.34	0.120	133, 219
	Physics coauthorship	Undirected	52 909	245 300	9.27	0.838	6.19	–	0.45	0.56	0.363	347, 349
	Biology coauthorship	Undirected	1 520 251	11 803 064	15.53	0.918	4.92	–	0.088	0.60	0.127	347, 349
	Telephone call graph	Undirected	47 000 000	80 000 000	3.16			2.1				10, 11
	Email messages	Directed	59 812	86 300	1.44	0.952	4.95	1.5/2.0		0.16		156
	Email address books	Directed	16 881	57 029	3.38	0.590	5.22	–	0.17	0.13	0.092	364
	Student dating	Undirected	573	477	1.66	0.503	16.01	–	0.005	0.001	−0.029	52
	Sexual contacts	Undirected	2 810					3.2				304, 305
Information	WWW `nd.edu`	Directed	269 504	1 497 135	5.55	1.000	11.27	2.1/2.4	0.11	0.29	−0.067	16, 41
	WWW AltaVista	Directed	203 549 046	1 466 000 000	7.20	0.914	16.18	2.1/2.7				84
	Citation network	Directed	783 339	6 716 198	8.57			3.0/–				404
	Roget's Thesaurus	Directed	1 022	5 103	4.99	0.977	4.87	–	0.13	0.15	0.157	272
	Word co-occurrence	Undirected	460 902	16 100 000	66.96	1.000		2.7		0.44		146, 175
Technological	Internet	Undirected	10 697	31 992	5.98	1.000	3.31	2.5	0.035	0.39	−0.189	102, 168
	Power grid	Undirected	4 941	6 594	2.67	1.000	18.99	–	0.10	0.080	0.003	466
	Train routes	Undirected	587	19 603	66.79	1.000	2.16	–		0.69	−0.033	425
	Software packages	Directed	1 439	1 723	1.20	0.998	2.42	1.6/1.4	0.070	0.082	−0.016	352
	Software classes	Directed	1 376	2 213	1.61	1.000	5.40	–	0.033	0.012	−0.119	453
	Electronic circuits	Undirected	24 097	53 248	4.34	1.000	11.05	3.0	0.010	0.030	−0.154	174
	Peer-to-peer network	Undirected	880	1 296	1.47	0.805	4.28	2.1	0.012	0.011	−0.366	6, 409
Biological	Metabolic network	Undirected	765	3 686	9.64	0.996	2.56	2.2	0.090	0.67	−0.240	252
	Protein interactions	Undirected	2 115	2 240	2.12	0.689	6.80	2.4	0.072	0.071	−0.156	250
	Marine food web	Directed	134	598	4.46	1.000	2.05	–	0.16	0.23	−0.263	245
	Freshwater food web	Directed	92	997	10.84	1.000	1.90	–	0.20	0.087	−0.326	321
	Neural network	Directed	307	2 359	7.68	0.967	3.97	–	0.18	0.28	−0.226	466, 470

Table 10.1: Basic statistics for a number of networks. The properties measured are: type of network, directed or undirected; total number of nodes n; total number of edges m; mean degree c; fraction of nodes in the largest component S (or the largest weakly connected component in the case of a directed network); mean distance between connected node pairs ℓ; exponent α of the degree distribution if the distribution follows a power law (or "–" if not; in/out-degree exponents are given for directed networks); clustering coefficient C from Eq. (7.28); clustering coefficient C_{WS} from the alternative definition of Eq. (7.31); and the degree correlation coefficient r from Eq. (7.64). The last column gives the citation(s) for each network in the References. Blank entries indicate unavailable data.

Componentes

- **Def:** Grafo simple no dirigido es conexo o conectado si cada vértice es alcanzable desde cualquier cualquier otro vértice (existe un camino)



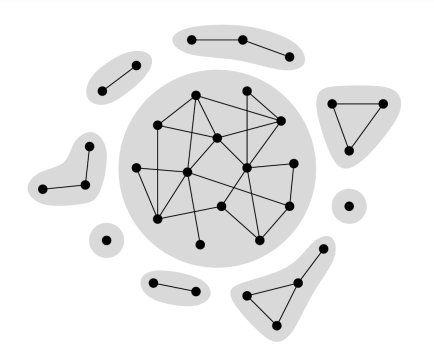
- **Def:** Una componente es un subgrafo conectado maximal
- La cantidad de componentes conexas es la multiplicidad del Laplaciano

Componente gigante

- En redes no dirigidas, típicamente existe una componente muy grande y el resto se divide en muchas componentes pequeñas desconectadas del resto

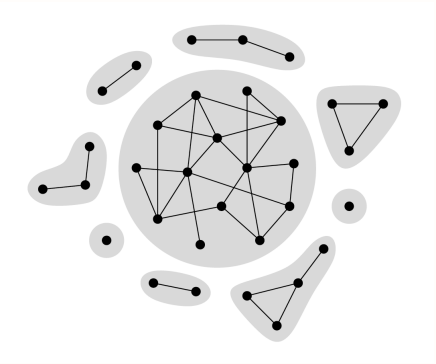
Componente gigante

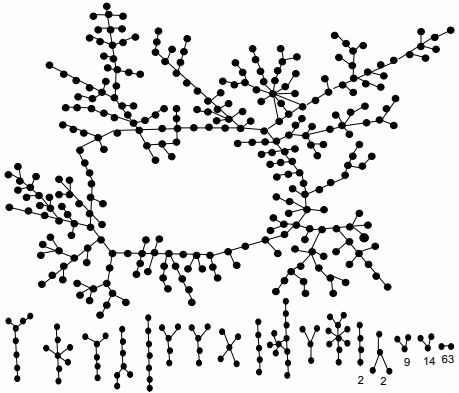
- En redes no dirigidas, típicamente existe una componente muy grande y el resto se divide en muchas componentes pequeñas desconectadas del resto



Componente gigante

- En redes no dirigidas, típicamente existe una componente muy grande y el resto se divide en muchas componentes pequeñas desconectadas del resto





- Aspecto típico y el de la red de relaciones románticas en liceales (Bearman)

Componente “gigante"

- ¿Cuál es el tamaño de la componente más grande en redes ?

Componente “gigante"

- ¿Cuál es el tamaño de la componente más grande en redes ?
- S fracción de vértices en la componente más grande
  - $\Rightarrow$ S es típicamente mayor que 0.9 (ver tabla)
  - $\Rightarrow$ $S = 1$ es posible: internet, redes descubiertas por exploración
- Ya vimos que no podemos tener dos componentes “grandes”

Componente “gigante"

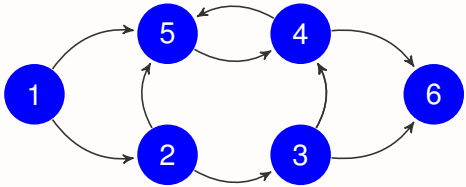
- ¿Cuál es el tamaño de la componente más grande en redes ?
- S fracción de vértices en la componente más grande
  - $\Rightarrow$ S es típicamente mayor que 0.9 (ver tabla)
  - $\Rightarrow$ $S = 1$ es posible: internet, redes descubiertas por exploración
- Ya vimos que no podemos tener dos componentes “grandes”
- ¿Se puede no tener ninguna componente grande? posible pero poco interesante para analizar con técnicas de network science...

Componentes en redes dirigidas

- El análisis de componentes en redes dirigidas es más complejo:

  $\Rightarrow$ Componente fuertemente conectada es equivalente al caso no dirigido

  $\Rightarrow$ Componente débilmente conectada si conectada al ignorar las direcciones de las aristas



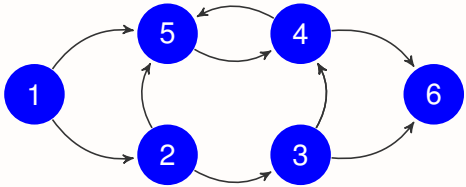
- Débilmente conexo pero no fuertemente conexo

Componentes en redes dirigidas

- El análisis de componentes en redes dirigidas es más complejo:

  $\Rightarrow$ Componente fuertemente conectada es equivalente al caso no dirigido

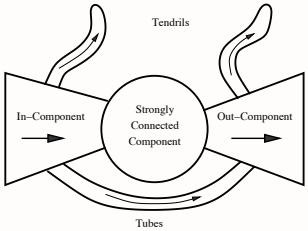
  $\Rightarrow$ Componente débilmente conectada si conectada al ignorar las direcciones de las aristas



- Débilmente conexo pero no fuertemente conexo
- Los grafos dirigidos acíclicos no tienen componentes fuertemente conexas (redes de citas de a 2 o 3)

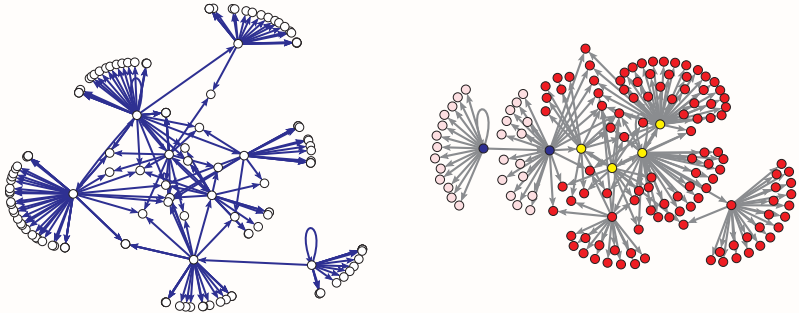
Estructura de “moña” de las redes dirigidas

- Primera aparición para la WWW en [Broder et al’00]



- El centro es la componente fuertemente conectada (SCC)
  - In-component (IC): vertices que alcanzan al centro, pero no viceversa
  - Out-component (OC): vertices que son alcanzados desde el centro, pero no viceversa
  - Tubos: vértices con conexiones entre la IC y la OC, pero no en SCC
  - Rizos: vértices que no se alcanzan ni son alcanzados por la SCC

Ejemplo: AIDS blog network



- Red de citas entre 146 blogs relacionadas al AIDS
  - $\Rightarrow$ SCC pequeña con 4 vértices e IC con 2 vértices
  - $\Rightarrow$ OC dominante con 112 vértices, y poco rizos (28 vértices)
- Para la WWW, Broder et al. encontraron que $|SCC| \approx |IC| \approx |OC|$ cada una un cuarto de la red (quizás desactualizado...)

Caminos más cortos y *small world*

- Las componentes gigantes tienden a mostrar la propiedad de small world
- Small refiere al largo de camino promedio

$$\bar{\ell} = \binom{N_v}{2}^{-1} \sum_{u \neq v \in V} d(u, v) = O(\log N_v)$$

Ex: facilita la dispersión de rumores, enfermedades, búsqueda de contenido en la WWW

Caminos más cortos y *small world*

- Las componentes gigantes tienden a mostrar la propiedad de small world
- Small refiere al largo de camino promedio

$$\bar{\ell} = \binom{N_v}{2}^{-1} \sum_{u \neq v \in V} d(u, v) = O(\log N_v)$$

Ex: facilita la dispersión de rumores, enfermedades, búsqueda de contenido en la WWW

- Watts y Strogatz (1998) encontraron que distancias cortas iban acompañadas de coefficiente de clustering alto y proponen un modelo

Caminos más cortos y *small world*

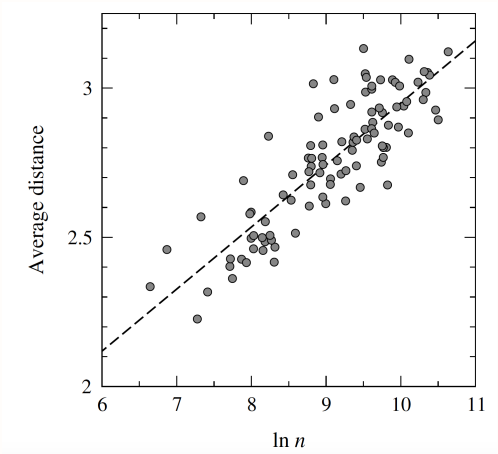
- Las componentes gigantes tienden a mostrar la propiedad de small world
- Small refiere al largo de camino promedio

$$\bar{\ell} = \binom{N_v}{2}^{-1} \sum_{u \neq v \in V} d(u, v) = O(\log N_v)$$

Ex: facilita la dispersión de rumores, enfermedades, búsqueda de contenido en la WWW

- Watts y Strogatz (1998) encontraron que distancias cortas iban acompañadas de coefficiente de clustering alto y proponen un modelo
- No es tan sorprendente la validez de la propiedad. Argumento informal:
- If $d_v \approx d$, luego de h_* saltos, se tiene $d^{h_*} \approx N_v \Rightarrow \bar{\ell} \approx h_* = O(\log N_v)$

Caminos más cortos y *small world*



- Largo de camino promedio en relaciones de amistad en Facebook para 100 estudiantes en función de $\log(N_v)$
- La recta es el mejor ajuste lineal

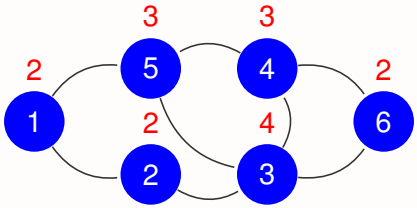
Diámetro de la red

- Se define el diámetro de una red como la máxima distancia entre dos vértices cualesquiera
  - $\Rightarrow$ longitud del camino más largo
  - $\Rightarrow$ también escala con $\log(N_v)$
- Poco robusto y un mal indicador de una red, pero sorprendentemente pequeño
- Efecto embudo (Milgram): muchos vértices son alcanzados a tráves de algunos pocos vértices de la red

Secuencia de grado

- **Def:** El grad d_v de v es el número de aristas incidentes

  $\Rightarrow$ La secuencia de grado ordena los grados en orden no-decreciente



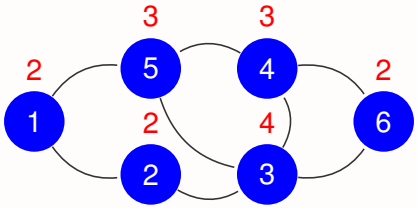
- En la figura $\Rightarrow$ los grados se muestran en rojo, e.g., $d_1 = 2$ y $d_5 = 3$

  $\Rightarrow$ la secuencia de grados es 2,2,2,3,3,4

Secuencia de grado

- **Def:** El grad d_v de v es el número de aristas incidentes

  $\Rightarrow$ La secuencia de grado ordena los grados en orden no-decreciente



- En la figura $\Rightarrow$ los grados se muestran en rojo, e.g., $d_1 = 2$ y $d_5 = 3$

  $\Rightarrow$ la secuencia de grados es 2,2,2,3,3,4

- En general, la secuencia de grados *no determina* el grafo

Distribución de grados

- Sea $N(d)$ el número de vértices con grado d

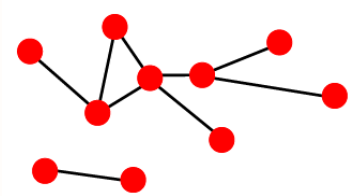
  $\Rightarrow$ La fracción de vértices con grado d e $\mathrm{P}(d) := \frac{N(d)}{N_v}$

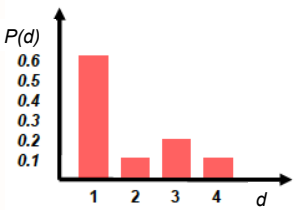
Distribución de grados

- Sea $N(d)$ el número de vértices con grado d

  $\Rightarrow$ La fracción de vértices con grado d e $\mathrm{P}(d) := \frac{N(d)}{N_v}$

- **Def:** la sucesión $\{\mathrm{P}(d)\}_{d \geq 0}$ es la distribución de grados de G
  - Podemos dibujar el histograma de la distribución de grado (bins de tamaño 1)

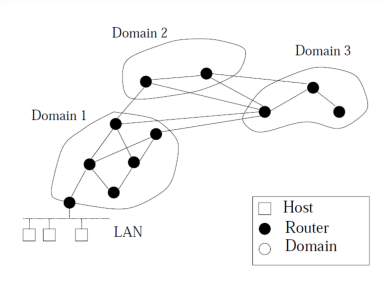


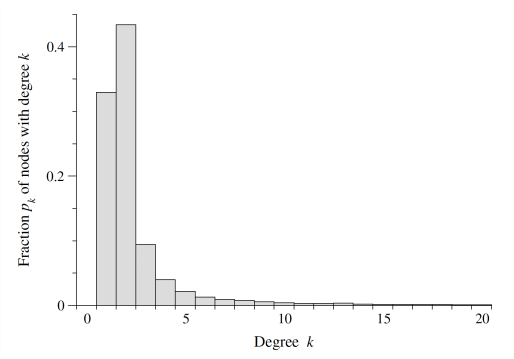


- $\mathrm{P}(d)$ es la probabilidad de que un vértice elegido al azar tenga grado d

Internet

- Topología de red de Internet [Faloutsos[3] '99]

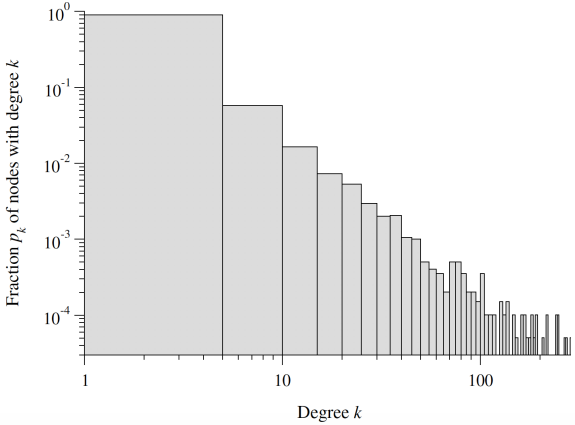


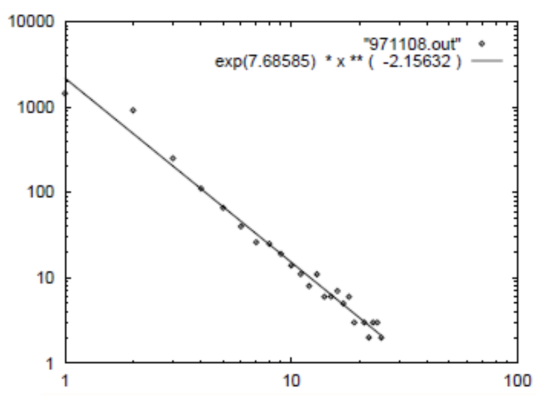


- Asimetría hacia la derecha también presente al nivel de routers y en otras redes

Internet

- Figura del Newman y ajuste de Faloutsos

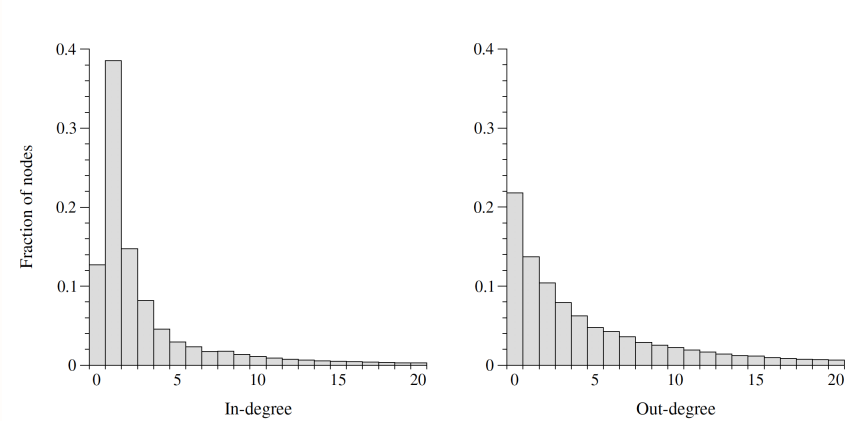




- Escala logarítmica y bins más grandes, se parece a una recta

World Wide Web

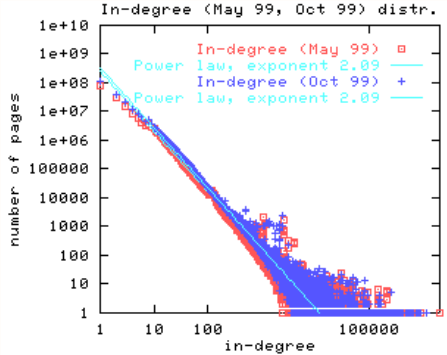
- Distribución de grado (entrante y saliente) de la WWW analizada en [Broder et al '00]

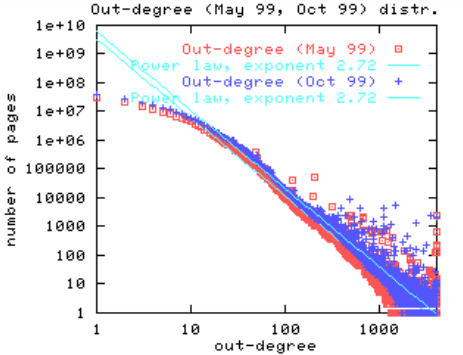


- Mayoría de los vértices tiene grado bajo

World Wide Web

- Cantidad no trivial de vértices con grado órdenes de magnitud más grande

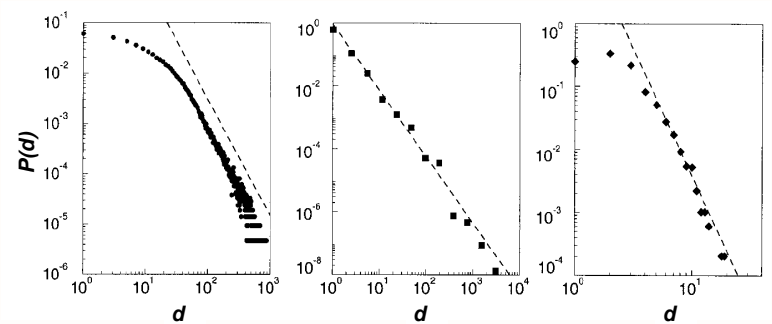




- Distibución de colas pesadas

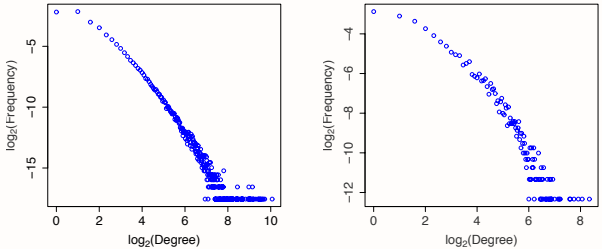
Figurita repetida

- Más distribuciones de colas pesadas se presentan en [Barabasi-Albert '99]



Author collaboration | Web graph | Power grid

Distribuciones de grado *power law*



- Figuras en escala log-log muestran un decaímiento aproximadamente lineal, que corresponde a una power law

$$\log \mathrm{P}(d) = -\alpha \log d + C \Rightarrow \mathrm{P}(d) \propto d^{-\alpha}$$

  - exponente *power-law* (menos la pendiente) usualmente está en $\alpha \in [2, 3]$
  - constante de normalización C no tiene mayores efectos

- *Power laws* suele representar mejor la cola de la distribución, i.e., for $d \geq d_{\text{mín}}$

Redes *Scale-free*

- Redes *scale-free*: redes cuya distribución de grados tiene colas *power law*
- ¿Porqué libres de escala? motivado en la propiedad de invariante por escala de las distribuciones *power law*

Redes *Scale-free*

- Redes *scale-free*: redes cuya distribución de grados tiene colas *power law*
- ¿Porqué libres de escala? motivado en la propiedad de invariante por escala de las distribuciones *power law*

Definición

Una función libre de escala $f(x)$ cumple que $f(ax) = bf(x)$, con $a, b \in \mathbb{R}$

Redes *Scale-free*

- Redes *scale-free*: redes cuya distribución de grados tiene colas *power law*
- ¿Porqué libres de escala? motivado en la propiedad de invariante por escala de las distribuciones *power law*

Definición

Una función libre de escala $f(x)$ cumple que $f(ax) = bf(x)$, con $a, b \in \mathbb{R}$

Ejemplo

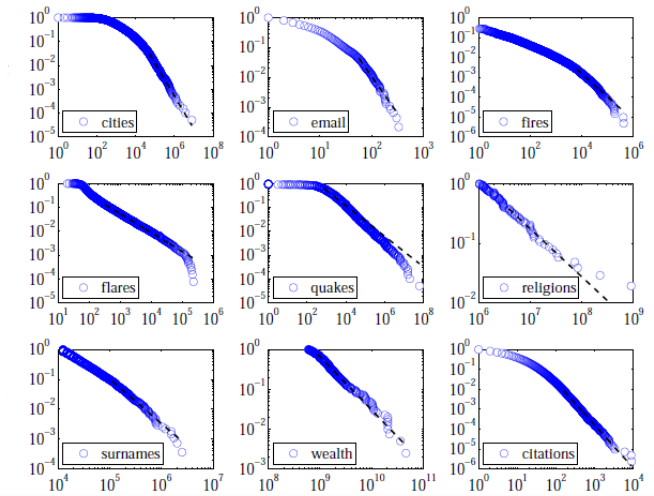
- Funciones power-law $f(x) = x^{-\alpha}$ son scale-free dado que

$$f(ax) = (ax)^{-\alpha} = a^{-\alpha} f(x) = bf(x), \text{ where } b := a^{-\alpha}$$

- Funciones exponencial $f(x) = c^x$ no son scale-free dado que

$$f(ax) = c^{ax} = (c^x)^a = f^a(x) \neq bf(x), \text{ except when } a = b = 1$$

Distribuciones power-law por doquier



- Clauset et al ’07 registraron distribuciones power law en muchos tipos de redes

Normalización

- Para ser distribución de probabilidad $\mathrm{P}(d) = Cd^{-\alpha}$ tiene que cumplirse que:

$$1 = \sum_{d=0}^{\infty} \mathrm{P}(d) = \sum_{d=0}^{\infty} Cd^{-\alpha} \;\Rightarrow\; C = \frac{1}{\sum_{d=0}^{\infty} d^{-\alpha}}$$

Normalización

- Para ser distribución de probabilidad $\mathrm{P}(d) = Cd^{-\alpha}$ tiene que cumplirse que:

$$1 = \sum_{d=0}^{\infty} \mathrm{P}(d) = \sum_{d=0}^{\infty} Cd^{-\alpha} \;\Rightarrow\; C = \frac{1}{\sum_{d=0}^{\infty} d^{-\alpha}}$$

- Si es válido solo para las colas $d \geq d_{\text{mín}}$, resulta que

$$C = \frac{1}{\sum_{d=d_{\text{mín}}}^{\infty} d^{-\alpha}} \approx \frac{1}{\int_{d_{\text{mín}}}^{\infty} x^{-\alpha} dx} = (\alpha - 1) d_{\text{mín}}^{\alpha - 1}$$

$\Rightarrow$ buena aproximación dado que $\mathrm{P}(d)$ varía lentamente con d

Definición

La distribución power law normalizada se define como

$$\mathrm{P}(d) = \frac{\alpha - 1}{d_{\text{mín}}} \left(\frac{d}{d_{\text{mín}}} \right)^{-\alpha}, \quad d \geq d_{\text{mín}}$$

Densidad *power law*

- Muchas veces es más sencillo pensar en distribuciones con densidad, $d \in \mathbb{R}_+$

Definición

Se define la densiad

$$p(d) = \frac{\alpha - 1}{d_{\text{mín}}} \left(\frac{d}{d_{\text{mín}}} \right)^{-\alpha}, \quad d \geq d_{\text{mín}}$$

- Es densidad, ya vimos que $\int_{d_{\text{mín}}}^{\infty} p(x)dx = 1$

  $\Rightarrow$ Para la convergencia de la integral tiene que ser $\alpha > 1$

Densidad *power law*

- Muchas veces es más sencillo pensar en distribuciones con densidad, $d \in \mathbb{R}_+$

Definición

Se define la densiad

$$p(d) = \frac{\alpha - 1}{d_{\text{mín}}} \left(\frac{d}{d_{\text{mín}}} \right)^{-\alpha}, \quad d \geq d_{\text{mín}}$$

- Es densidad, ya vimos que $\int_{d_{\text{mín}}}^{\infty} p(x)dx = 1$
  $\Rightarrow$ Para la convergencia de la integral tiene que ser $\alpha > 1$
- Ejemplo: Probabilidad de que un vértice al azar tenga grado mayor que 100 está dado por:

$$\mathrm{P}\left(D_v > 100\right) = \int_{100}^{\infty} \frac{\alpha - 1}{d_{\text{mín}}} \left(\frac{x}{d_{\text{mín}}} \right)^{-\alpha} dx = \left(\frac{100}{d_{\text{mín}}} \right)^{1-\alpha}$$

Momentos de una densidad *power law*

- ¿Cuál es el momento m-ésimo de una v.a. con distribución *power law*?
- Usando la definición resulta que:

$$\mathbb{E}\left[D_v^m\right] = \int_{d_{\text{mín}}}^{\infty} x^m p(x)dx = \frac{\alpha - 1}{d_{\text{mín}}^{1-\alpha}} \left[\frac{x^{m+1-\alpha}}{m+1-\alpha}\right]_{d_{\text{mín}}}^{\infty}$$

$\Rightarrow$ la convergencia del la integral requiere $m + 1 < \alpha$

Momentos de una densidad *power law*

- ¿Cuál es el momento m-ésimo de una v.a. con distribución *power law*?
- Usando la definición resulta que:

$$\mathbb{E}[D_v^m] = \int_{d_{\text{mín}}}^{\infty} x^m p(x) dx = \frac{\alpha - 1}{d_{\text{mín}}^{1-\alpha}} \left[\frac{x^{m+1-\alpha}}{m+1-\alpha} \right]_{d_{\text{mín}}}^{\infty}$$

$\Rightarrow$ la convergencia del la integral requiere $m + 1 < \alpha$

- Típicamente se observa $\alpha \in (2, 3)$ de donde

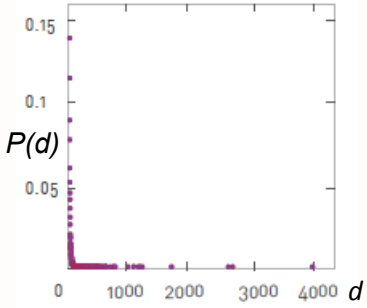
$$\mathbb{E}[D_v] = \left(\frac{\alpha - 1}{\alpha - 2} \right) d_{\text{mín}} < \infty \text{ y } \mathbb{E}[D_v^m] = \infty,\ m \geq 2$$

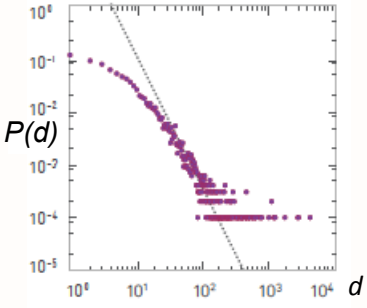
- En particualr el segundo momento y la varianza valen infinito

$\Rightarrow$ la desviación éstandar es una escala para las distribuciones con escala

Visualizando distribuciones *power law*

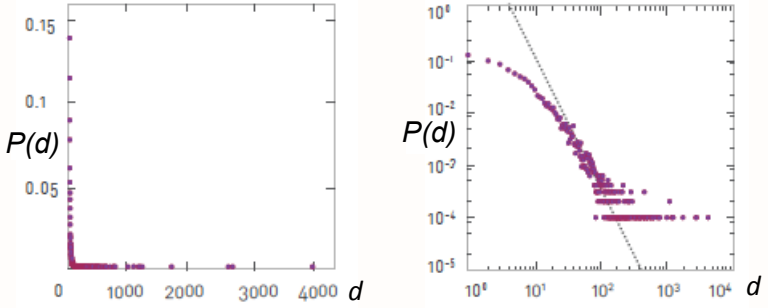
- Se observa en escala log-log: acumula probabilidades y expande grados





Visualizando distribuciones *power law*

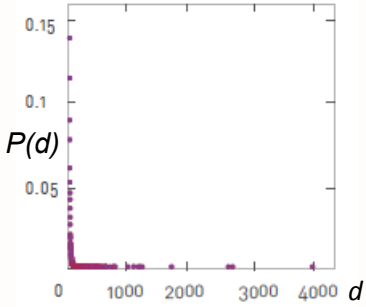
- Se observa en escala log-log: acumula probabilidades y expande grados

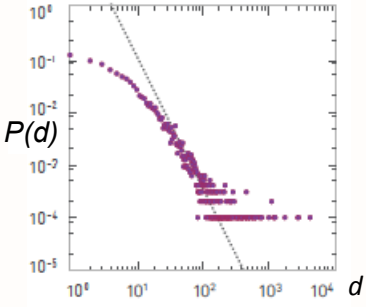


- Con particiones de largo 1 no hay suficiente resolución para los grados altos

Visualizando distribuciones *power law*

- Se observa en escala log-log: acumula probabilidades y expande grados





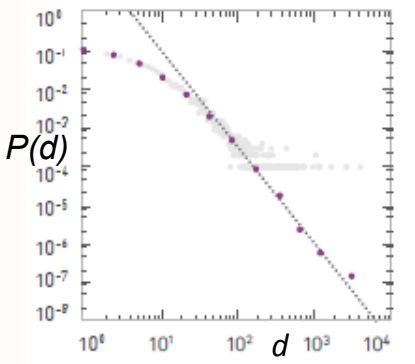
- Con particiones de largo 1 no hay suficiente resolución para los grados altos
- Se usan particiones logarítmicas de la forma

$$a^{n-1} \leq d < a^n, \quad n = 1, 2, \ldots$$

Partición logarítmica

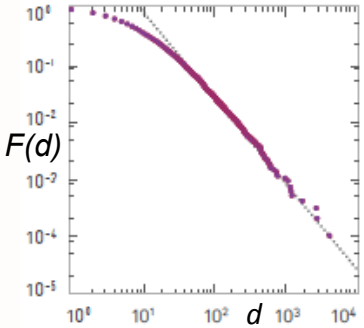
- Ejemplo: si $a = 2$, el n-ésimo intervalo tiene largo $2^n - 2^{n-1} = 2^{n-1}$.

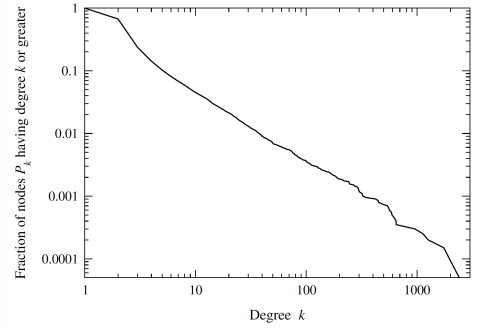
  $\Rightarrow$ se normaliza por el conteo por el largo del intervalo



Función de distribución complementaria

- La función de distribución complmeentaria (CCDF) es $\bar{F}(d) = \mathrm{P}(D_v \geq d)$
- Si la densidad es power law con densidad α, la CCDF también es power law con exponente $\alpha - 1$
- Dibujar la CCDF en escala log-log y buscar ajuste a una recta





Estimación del exponente

- No es buena idea hallarlo a partir de las pendientes de las rectas (Mínimos Cuadrados)

  $\Rightarrow$ para la densidad, el logaritmo distorsiona los errores de manera diferente y hay que elegir el d_{min}

  $\Rightarrow$ para la CCDF puntos consecutivos son dependientes

- Estimador de máxima verosimilitud (EMV)

$$\hat{\alpha} = 1 + \left[\frac{1}{N_v} \sum_{i=1}^{N_v} \log \left(\frac{d_i}{d_{\text{mín}}} \right) \right]^{-1}$$

  $\Rightarrow$ hay que elegir el d_{min} !

Estimación del exponente

- No es buena idea hallarlo a partir de las pendientes de las rectas (Mínimos Cuadrados)

  $\Rightarrow$ para la densidad, el logaritmo distorsiona los errores de manera diferente y hay que elegir el d_{min}

  $\Rightarrow$ para la CCDF puntos consecutivos son dependientes

- Estimador de máxima verosimilitud (EMV)

$$\hat{\alpha} = 1 + \left[\frac{1}{N_v} \sum_{i=1}^{N_v} \log \left(\frac{d_i}{d_{\text{mín}}} \right) \right]^{-1}$$

  $\Rightarrow$ hay que elegir el d_{min} !

  $\Rightarrow$ a mano o Hill plot...

Hill plot del EMV

1) Obtener la secuencia de grados $d_{(1)} \leq \ldots \leq d_{(N_v)}$

Hill plot del EMV

1) Obtener la secuencia de grados $d_{(1)} \leq \ldots \leq d_{(N_v)}$

2) Para cada $k \in \{1, \ldots, N_v - 1\}$ sea $d_{\text{mín}} = d_{(N_v - k)}$. El EMV es

$$\hat{\alpha}(k) = 1 + \left[\frac{1}{k} \sum_{i=0}^{k-1} \log \left(\frac{d_{(N_v - i)}}{d_{(N_v - k)}} \right) \right]^{-1}$$

Hill plot del EMV

1) Obtener la secuencia de grados $d_{(1)} \leq \ldots \leq d_{(N_v)}$

2) Para cada $k \in \{1, \ldots, N_v - 1\}$ sea $d_{\text{mín}} = d_{(N_v - k)}$. El EMV es

$$\hat{\alpha}(k) = 1 + \left[\frac{1}{k} \sum_{i=0}^{k-1} \log \left(\frac{d_{(N_v - i)}}{d_{(N_v - k)}} \right) \right]^{-1}$$

3) Dibujar y examinar el Hill plot de $\hat{\alpha}(k)$ en función de k

➤ Si una distribución *power law* es creíble, el Hill plot debería 'estabilizarse'

$\Rightarrow$ Identificar $\hat{\alpha}$ para un rango amplio de valores intermedios de k

Hill plot del EMV

1) Obtener la secuencia de grados $d_{(1)} \leq \ldots \leq d_{(N_v)}$

2) Para cada $k \in \{1, \ldots, N_v - 1\}$ sea $d_{\text{mín}} = d_{(N_v - k)}$. El EMV es

$$\hat{\alpha}(k) = 1 + \left[\frac{1}{k} \sum_{i=0}^{k-1} \log \left(\frac{d_{(N_v - i)}}{d_{(N_v - k)}} \right) \right]^{-1}$$

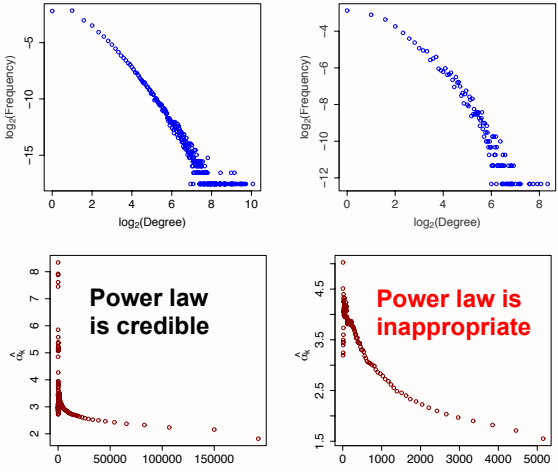
3) Dibujar y examinar el Hill plot de $\hat{\alpha}(k)$ en función de k

- Si una distribución *power law* es creíble, el Hill plot debería 'estabilizarse'

  $\Rightarrow$ Identificar $\hat{\alpha}$ para un rango amplio de valores intermedios de k

- ¿Porqué valores intermedios?
  - k chicos: estimación poco exacta porque hay pocos datos
  - k grandes: sesgos si la *power law* es válida solo en la cola

Ejemplo: Internet e interacción de proteínas



- Un decaimiento brusco en $\hat{\alpha}_k$ suggiere que un modelo simple de *power law* no es apropiado

Distribución de otras medidas de centralidad

- El grado es también una medida de centralidad de vector propio

Distribución de otras medidas de centralidad

- El grado es también una medida de centralidad de vector propio
- Es de esperar que las otras medidas de centralidad de vector propio también presenten una distribución *power law*

Distribución de otras medidas de centralidad

- El grado es también una medida de centralidad de vector propio
- Es de esperar que las otras medidas de centralidad de vector propio también presenten una distribución *power law*
- La betweenness también presenta asimetrías hacia la derecha y también puede decirse que en muchos casos presenta distibución *power law*

Distribución de otras medidas de centralidad

- El grado es también una medida de centralidad de vector propio
- Es de esperar que las otras medidas de centralidad de vector propio también presenten una distribución *power law*
- La betweenness también presenta asimetrías hacia la derecha y también puede decirse que en muchos casos presenta distibución *power law*
- No es así para la centralidad de cercanía, típicamente tiene un rango de variación acotado

  $\Rightarrow$ el valor máximo es el diámetro de la red y ya vimos que tienden a ser valores chicos

Coeficiente de Clustering

- En la tabla se observan valores de entre 0.1 y 0.6 (densidad de triángulos) que resultan ser valores altos

Coeficiente de Clustering

- En la tabla se observan valores de entre 0.1 y 0.6 (densidad de triángulos) que resultan ser valores altos
- Se puede probar que para una secuencia de grados dada, si las aristas se eligen al azar, el coeficiente de clustering está dado por:

$$C = \frac{1}{n} \frac{\left(\langle k^2 \rangle - \langle k \rangle\right)^2}{\langle k^3 \rangle}$$

donde $\langle k^m \rangle = \sum_k k^m p(k)$ es el m-ésimo momento.

Coeficiente de Clustering

- En la tabla se observan valores de entre 0.1 y 0.6 (densidad de triángulos) que resultan ser valores altos
- Se puede probar que para una secuencia de grados dada, si las aristas se eligen al azar, el coeficiente de clustering está dado por:

$$C = \frac{1}{n} \frac{\left(\langle k^2 \rangle - \langle k \rangle\right)^2}{\langle k^3 \rangle}$$

donde $\langle k^m \rangle = \sum_k k^m p(k)$ es el m-ésimo momento.

- Si la distribución no cambia, al crecer n, se tiene que $c \to 0$.

Coeficiente de Clustering

- En la tabla se observan valores de entre 0.1 y 0.6 (densidad de triángulos) que resultan ser valores altos
- Se puede probar que para una secuencia de grados dada, si las aristas se eligen al azar, el coeficiente de clustering está dado por:

$$C = \frac{1}{n} \frac{\left(\langle k^2 \rangle - \langle k \rangle\right)^2}{\langle k^3 \rangle}$$

donde $\langle k^m \rangle = \sum_k k^m p(k)$ es el m-ésimo momento.

- Si la distribución no cambia, al crecer n, se tiene que $c \to 0$.
- Los valores de la tabla son mayores a los que se obtienen con la fórmula
  - $\Rightarrow$ las relaciones no son al azar, y aparece la estructura
  - $\Rightarrow$ dependiendo del contexto se favorecen o se evitan las creaciones de triángulos

(dis)Assortative mixing

- En la tabla se muestra el valor del coeficiente de correlación r para la asortatividad por grado pause
- Es difícil de calcular como lo definimos antes. Se usa la siguiente expresión:

$$r = \frac{S_1 S_e - S_2^2}{S_1 S_3 - S_2^2} \quad \text{donde } S_i = \sum_{v \in V} d_v^i \text{ y } S_e = \sum_{ij} A_{ij} d_i d_j = 2 \sum_{(i,j) \in E} d_i d_j$$

(dis)Assortative mixing

- En la tabla se muestra el valor del coeficiente de correlación r para la asortatividad por grado pause
- Es difícil de calcular como lo definimos antes. Se usa la siguiente expresión:

$$r = \frac{S_1 S_e - S_2^2}{S_1 S_3 - S_2^2} \quad \text{donde } S_i = \sum_{v \in V} d_v^i \text{ y } S_e = \sum_{ij} A_{ij} d_i d_j = 2 \sum_{(i,j) \in E} d_i d_j$$

- Se observa que las redes sociales tienen $r > 0$ (assortative mixing por grado) mientras que el resto tiene $r < 0$ (disassortative mixing por grado)

(dis)Assortative mixing

- En la tabla se muestra el valor del coeficiente de correlación r para la asortatividad por grado pause
- Es difícil de calcular como lo definimos antes. Se usa la siguiente expresión:

$$r = \frac{S_1 S_e - S_2^2}{S_1 S_3 - S_2^2} \quad \text{donde } S_i = \sum_{v \in V} d_v^i \text{ y } S_e = \sum_{ij} A_{ij} d_i d_j = 2 \sum_{(i,j) \in E} d_i d_j$$

- Se observa que las redes sociales tienen $r > 0$ (assortative mixing por grado) mientras que el resto tiene $r < 0$ (disassortative mixing por grado)

  $\Rightarrow$ difícil que en redes simples se junten muchos vértices de grados altos con vértices de grados altos

  $\Rightarrow$ en redes sociales se forman grupos y entre los grupos los grados se reducen