Análisis de Redes Sociales

Guillermo Jiménez Díaz (gjimenez@ucm.es)
Alberto Díaz (albertodiaz@fdi.ucm.es)

22 de octubre de 2014

Prefacio

Estos son los apuntes de la asignatura Análisis de Redes Sociales, impartida en la Facultad de Informática de la Universidad Complutense de Madrid por los profesores Guillermo Jiménez Díaz y Alberto Díaz, del Departamento de Ingeniería del Software e Inteligencia Artificial.

Este material ha sido desarrollado a partir de distintas fuertes, destacando como referencia principal el libro *Network Science* de Laszlo Barabasi, el material de la asignatura *Social Network Analysis*, impartido por Lada Adamic a través de Coursera, y las transparencias de la asignatura Redes y Sistemas Complejos, creadas por Oscar Cordón García de la Universidad de Granada.

Tema 3: Redes libres de escala

Tal y como hemos visto en el capítulo anterior el modelo de Erdös-Renyi o de red aleatoria presenta algunas propiedades que no se cumplen en las redes reales. De entre ellas destacamos que el modelo de redes aleatoria no contempla la existencia de hubs o concentradores, nodos con un grado mucho mayor que el resto de nodos de la red.

En este tema vamos a estudiar un nuevo modelo de red que se ajusta mucho mejor a las redes reales, conocido como **redes libres de escala**.

3.1 La WWW: el primer mapa de red no aleatoria

En 1999 un grupo de investigadores realizó el primer mapa de una porción de la WWW con el objetivo de verificar las propiedades de los modelos de red aleatoria en este sistema complejo de páginas y enlaces. Hasta aquel momento había razones suficientes para suponer que esta red seguía este modelo de red aleatoria. Compusieron el mapa del dominio `nd.edu`, con un total de 300000 documentos (nodos) y 1,5 millones de enlaces.

Aunque a primera vista parecía que este mapa cumplía esa aleatoriedad predicha por el modelo. Sin embargo, se pudo observar que, conviviendo con un gran número de nodos de grado pequeño, existen nodos con un número excepcional de enlaces: los hubs o concentradores. Recordemos que estos nodos son extremadamente improbables en las redes aleatorias por lo que este estudio dio lugar a replantearse que tal vez este modelo no era el que servía para modelar esta red. Así mismo, como veremos más adelante, se observó que estos hubs aparecen en otras muchas redes reales.

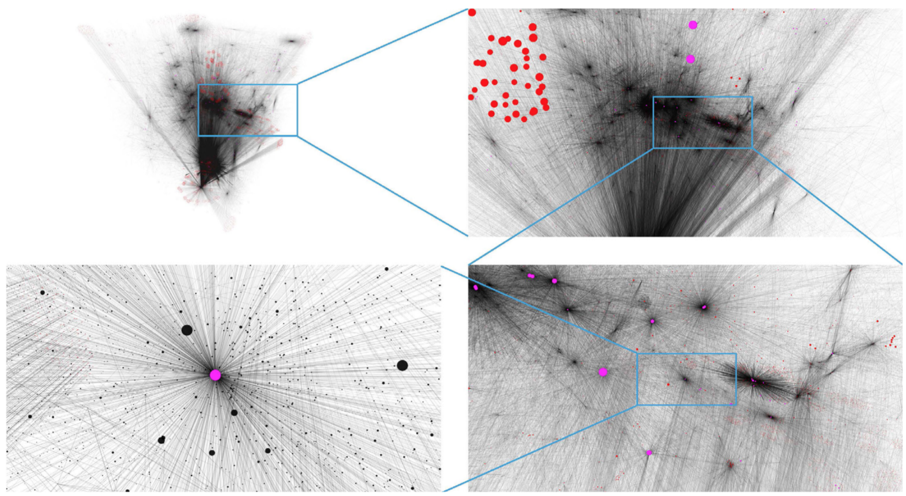

Figure 1: Mapa de la WWW: se puede ver la existencia de hubs.

El análisis de los datos de la WWW muestra que la distribución de grados no se asemeja a la distribución de Poisson que predice el modelo de redes aleatorias sino que dicha distribución se asemeja a una **ley potencial** o *power law*. Esta distribución se caracteriza por tener una cola mucho más ancha que una exponencial. La diferencia entre una y otra es mucho más clara si las representamos en una escala logarítmica.

De acuerdo a esto, la distribución de grados en la WWW se puede representar como:

$$p_k \sim k^{-\gamma} \rightarrow log\, p_k \sim -\gamma \cdot log\, k$$

Como la WWW es una red dirigida podemos distinguir entre la distribución de grados de entrada y la distribución de grados de salida:

$$p_{k_{in}} \sim k^{-\gamma_{in}} \quad p_{k_{out}} \sim k^{-\gamma_{out}}$$

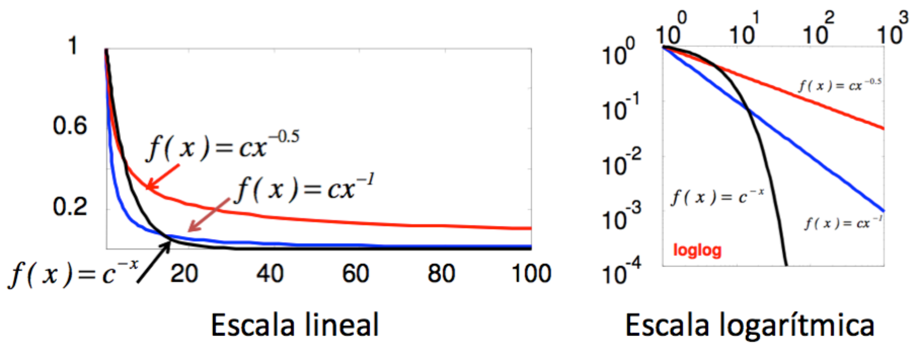

Figure 2: Comparativa de una función exponencial (negro) frente a funciones de ley potencial (azul y rojo). A la izquierda se presenta en escala lineal mientras que a la derecha se ve en escala logarítmica

De acuerdo a los datos usados por Barabasi y Albert se pudo estimar que $\gamma_{in} \approx 2,1$ mientras que $\gamma_{in} \approx 2,45$.

Lo realmente interesante de este estudio no fueron los datos resultantes obtenidos sino la gran diferencia entre la distribución esperada (una Poisson) y la distribución final obtenida (una ley potencial). En la siguiente figura se observa claramente estas diferencias obtenidas:

Esto dio a entender que esta red real no se comportaba de acuerdo al modelo de red aleatoria sino que era "otro tipo de red"

3.2 Redes libres de escala

Una **red libre de escala** es una red cuya distribución de grados sigue una ley potencial. Esta distribución se caracteriza por tener una larga cola. Para esta distribución se pueden emplear dos formulaciones distintas aunque las propiedades de la red son independientes del formalismo empleado.

3.2.1 Formalismo discreto

En este caso consideramos que el grado de un nodo es siempre un entero positivo ($K = 0, 1, 2 \ldots$). De acuerdo a esto, la probabilidad de que un nodo aleatorio tenga grado k se puede expresar como:

$$p_k = C \cdot k^{-\gamma}$$

La constante C se puede calcular teniendo en cuenta que $\sum_{k=1}^{\infty} p_k = 1$:

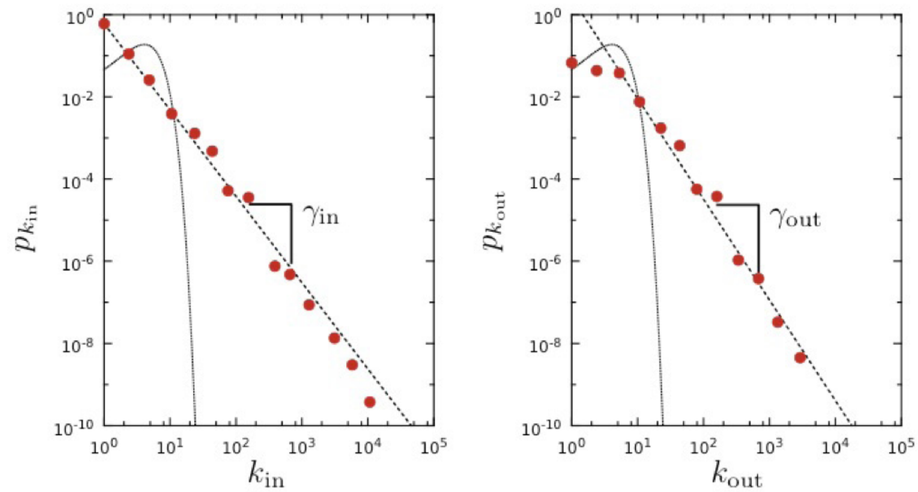

Figure 3: Distribución de grados de entrada y de salida de la WWW frente a la Poisson (en escala logarítmica)

$$C = \frac{1}{\sum_{k=1}^{\infty} k^{-\gamma}} = \frac{1}{\varsigma(\gamma)}$$

donde $\varsigma(\gamma)$ es la función Zeta de Riemann.

Para valores de $k > 0$ se cumple que la función de distribución es:

$$p_k = \frac{k^{-\gamma}}{\varsigma(\gamma)}$$

Los nodos aislado ($k = 0$) tienen que tratarse de manera independiente.

3.2.2 Formalismo contínuo

En este formalismo consideramos que k puede tomar cualquier valor real positivo. En este caso:

$$p(k) = C \cdot k^{-\gamma}$$

La constante C se puede calcular teniendo en cuenta que $\int_{k_{min}}^{\infty} p(k)dk = 1$:

$$C = \frac{1}{\int_{k_{min}}^{\infty} k^{-\gamma} dk} = (\gamma - 1) \cdot k_{min}^{\gamma-1}$$

De acuerdo a esto, la función de distribución según el formalismo continuo es:

$$p(k) = (\gamma - 1) \cdot k_{min}^{\gamma-1} \cdot k^{\gamma-1}$$

Donde k_{min} es el menor grado de la red (o función de distribución). En este caso, la función de distribución se ha de interpretar como la probabilidad de que un nodo tenga un grado entre dos valores k_1 y k_2:

$$\int_{k_1}^{k_2} p(k)dk$$

El primer informe relacionado con una función de ley potencial fue realizado por Pareto (conocido por la ley 80/20), donde destacaba que durante el siglo XIX, el 80% del dinero en Italia estaba en manos de solo un 20% de la población.

3.3 Hubs en las redes libres de escala

La principal diferencia entre la función de distribución en una red aleatoria frente a la de una red libre de escala es la cola de la distribución, observándose más claramente en la escala logarítmica. La principal implicación de esta cola es que las redes libres de escala contemplan la existencia de **hubs**, algo poco probable en una red aleatoria.

Analizando las distribuciones de grados de ambas redes más en detalle podemos ver que:

- La distribución de grados de la red libre de escala presenta una probabilidad de la existencia de nodos con un k pequeño mayor que en una red aleatoria (donde la probabilidad de estos nodos es casi nula).
- La red aleatoria tiene una mayor probabilidad de tener nodos con un grado $k \sim \langle k \rangle$ que la red libre de escala (ya que la función de distribución de la primera está por encima de la segunda)
- La red libre de escala presenta una mayor probabilidad de que existan nodos con k muy alto. De hecho, la probabilidad de que existan hubs en una red libre de escala es varios órdenes de magnitud mayor que en una red aleatoria (aunque este hecho no se aprecia en la escala lineal podemos verlo claramente reflejado en la escala logarítmica)

A modo de ejemplo, si usamos los datos de la WWW y suponemos que es una red aleatoria (donde hemos calculado que $\langle k \rangle = 4,6$) la probabilidad de que un nodo tenga 100 enlaces es $p_{aleatoria}(100) = 10^{-30}$. Sin embargo, suponiendo que es una red libre de escala esta probabilidad desciende a $p_{libre-escala}(100) = 10^{-4}$.

Para cualquier p_k podemos calcular el máximo grado esperado k_{max}, es decir, el grado del mayor hub existente en la red:

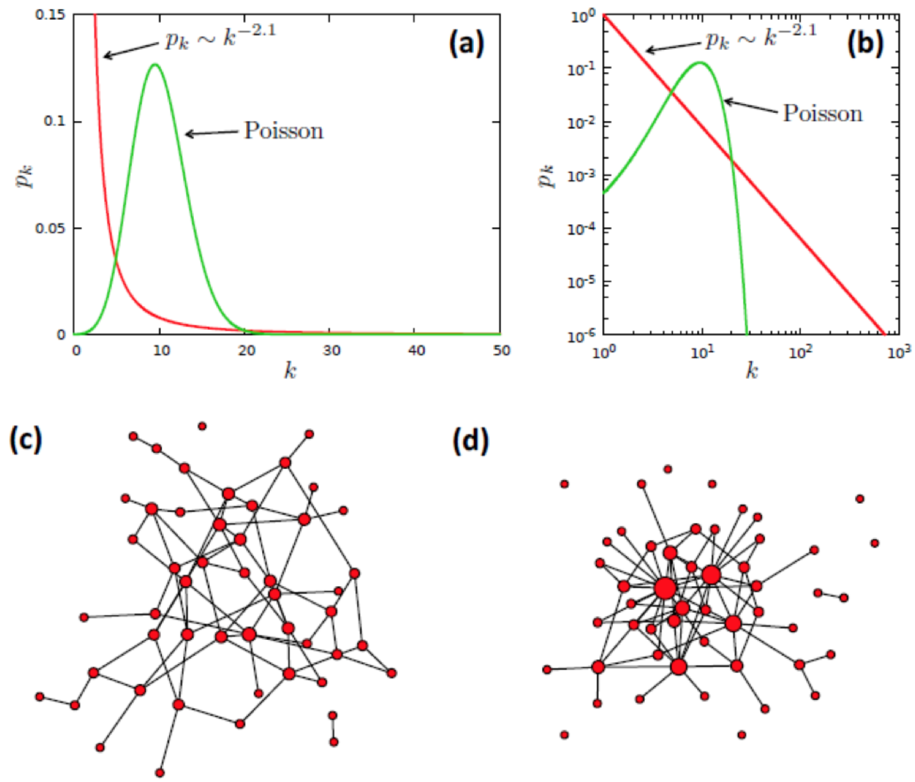

Figure 4: Distribuciones para redes aleatorias y redes libres de escala ((a) en escala lineal; (b) en escala logarítmica y representación gráfica de una (c)red aleatoria frente a una (d)red libre de escala ($\langle k \rangle = 3$; $N = 50$)

- Si suponemos que tenemos una red aleatoria con una distribución de grados p_k exponencial entonces tenemos que:

$$p_k = C \cdot e^{-\lambda k} : \quad k_{max} = k_{min} + \frac{lnN}{\lambda}$$

El tamaño del hub más grande (k_{max}) depende de lnN, que es una función que crece suavemente, por lo que hay muy poca diferencia entre k_{min} y k_{max}. Esto implica que no hay hubs.

- Si suponemos que tenemos una red aleatoria con una distribución de grados p_k que sigue una Poisson entonces tenemos que la dependencia de k_{max} es aún más suave por lo que hay aún menor diferencia entre k_{min} y k_{max}.

- Si suponemos que tenemos una red libre de escala y, por tanto, con una distribución de grados p_k de ley potencial entonces tenemos que:

$$p(k) = C \cdot k^{-\gamma} \quad k_{max} \sim k_{min} \cdot N^{\frac{1}{\gamma-1}}$$

En este caso, el tamaño del hub mayor depende de N. Esto implica que cuanto mayor es la red mayor es el grado del hub más grande.

De nuevo, si tomamos los datos de una muestra de la WWW con $N = 3 \cdot 10^5$ nodos:

$$k_{max} \approx 13 \quad \text{si la red es aleatoria}$$

$$k_{max} \approx 85000 \quad \text{si la red es libre de escala}$$

Podemos ver gráficamente esta dependencia:

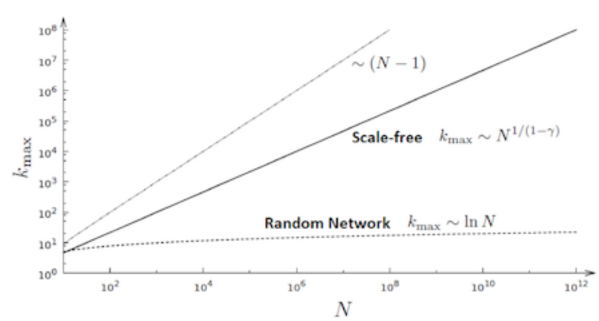

Figure 5: Relación entre el grado del mayor hub (k_{max}) y el tamaño de la red (N)

A modo de resumen podemos decir que las redes aleatorias están “acotadas”, en el sentido de que todos sus nodos tienen un grado similar y la existencia de hubs es muy improbable. En cambio, las redes libres de escala no están “acotadas” (son libres, de ahí su nombre como veremos en la próxima sección) y se espera que existan hubs junto con nodos de grado muy pequeño. Además hemos visto que el tamaño de la red influye en el tamaño de los hubs. Más gráficamente podemos comparar la red de carreteras de EEUU (una red aleatoria) frente a la red de tráfico aéreo del mismo país (red libre de escala). Como veremos más adelante, el tipo de red afecta a las distancias y a la forma en la que se navega por ellas.

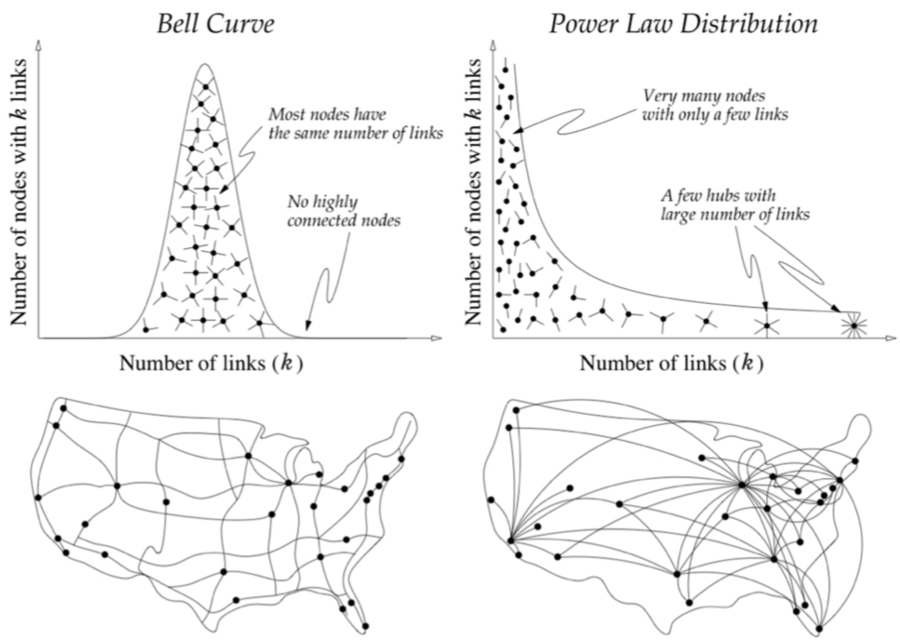

Figure 6: Red de carreteras de EEUU (una red aleatoria) frente a la red de tráfico aéreo del mismo país (red libre de escala)

3.3.1 Origen del término

El término “libre de escala” está relacionado con los momentos de la distribución de probabilidad de grados.

$$\langle k^n \rangle = \sum_{k_{min}}^{\infty} k^n p_k = \int_{k_{min}}^{\infty} k^n p(k) dk$$

Recordemos que:

- n=1: El primer momento es el grado medio ($\langle k \rangle$).
- n=2: el segundo momento es la varianza ($\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2$), la cual mide la dispersión de los grados. su raíz cuadrada es la desviación típica (σ)
- n=3: el tercer momento $\langle k^3 \rangle$ determina la asimetría, es decir, cómo de simétrica es p_k alrededor de la media (si la función es simétrica entonces $\langle k^3 \rangle = 0$)

En una red libre de escala el momento n-ésimo es:

$$\langle k^n \rangle = \int_{k_{min}}^{k_{max}} k^n p(k) dk = C \frac{k_{max}^{n-\gamma+1} - k_{min}^{n-\gamma+1}}{n-\gamma+1}$$

Como ya hemos deducido anteriormente, k_{max} crece con el tamaño de la red por lo que analizaremos su comportamiento en el límite $k_{max} \to \infty$:

$$\langle k^n \rangle = \frac{(\gamma-1)}{(n-\gamma-1)} k_{min}^{\gamma-1} \Big[k^{n-\gamma+1} \Big]_{k_{min}}^{\infty}$$

De aquí se deduce que:

- Si $n - \gamma + 1 \leq 0$ entonces el momento está acotado, por lo que todos los momentos en los que $n \leq \gamma + 1$ son finitos.
- Si $n - \gamma + 1 > 0$ entonces el momento tiende a infinito según la red crece, por lo que todos los momentos en los que $n > \gamma + 1$ divergen.

El uso de los datos de distintas redes reales muestra que el exponente del grado γ tiene un valor entre 2 y 3. Según lo anterior se deduce que el primer momento ($\langle k \rangle$) es finito pero que la varianza (segundo momento o $\langle k^2 \rangle$) y el resto de momentos tienden a infinito. Si esto es así entonces resulta que los valores medios no tienen realmente sentido ya que hay fluctuaciones demasiado grandes. La escala interna (la media) no tiene sentido por lo que es una red "libre de escala".

Más adelante estudiaremos más en detalle la implicación del valor del exponente γ.

3.4 Universalidad en las redes libres de escala

Finalmente, podemos ver que las redes reales cumplen otra nueva ley que es la **propiedad libre de escala**:

> Muchas redes reales presentan distribuciones de cola ancha. Esto implica que nodos de grado bajo conviven con nodos con un grado excepcionalmente grande: los hubs.

NETWORK	NL		$\langle k \rangle$ $\langle k_{in} \rangle = \langle k_{out} \rangle$	σ_{in}	σ_{out}	σ	γ_{in}	γ_{out}	γ
Internet	192,244	609,066	6.34	-	-	14.14	-	-	3.42*
WWW	325,729	1,497,134	4.60	39.05	21.48	-	2.31	2.00	-
Power Grid	4,941	6,594	2.67	-	-	1.79	-	-	Exp.
Mobile Phone Calls	36,595	91,826	2.51	2.39	2.32	-	4.69*	5.01*	-
Email	57,194	103,731	1.81	9.56	34.07	-	3.43*	2.03	-
Science Collaboration	23,133	93,439	8.08	-	-	10.63	-	-	3.35
Actor Network	702,388	29,397,908	83.71	-	-	200.86	-	-	2.12
Citation Network	449,673	4,689,479	10.43	29.37	9.49	-	3.03**	4.00	-
E. Coli Metabolism	1,039	5,802	5.58	22.46	19.12	-	2.43	2.90	-
Yeast Protein Interactions	2,018	2,930	2.90	-	-	4.88	-	-	2.89*

Figure 7: Datos de redes reales. Podemos ver que la mayoría de los exponentes γ tienen un valor entre 2 y 3

At a glance: Random networks

- *Definition*: N nodes, where each node pair is connected with probability p.

- *Average degree:* $\langle k \rangle = p(N-1)$

- *Average number of links:* $\langle L \rangle = \frac{p(N-1)}{2}$

- *Degree distribution:* $p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}$.

For sparse networks *($k \ll N$)*, P_k has the Poisson form

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}.$$

- *Giant component* (N_G):

$\langle k \rangle < 1$: no giant component $(N_G \sim \ln N)$

$1 < \langle k \rangle < \ln N$: one giant component and disconnected clusters

$$\left(N_G \sim N^{\frac{2}{3}} \right)$$

$\langle k \rangle > \ln N$: all nodes join the giant component $N_G \sim (p - p_c) N$

- *Average distance:* $\langle d \rangle \propto \frac{\log N}{\log \langle k \rangle}$,

- *Clustering coefficient:* $C = \frac{\langle k \rangle}{N}$.

Box 3.8

Figure 8: Cuadro resumen (extraído de *Network Science*, pp. 70)