Universidad de Buenos Aires

Informe 1

Redes Complejas

Raúl Barriga
Mariela Celis
Jimmy Masías
Sebastían Pinto

15 de septiembre de 2016

Índice

1. Ejercicio 1 **1**

2. Ejercicio 2. **2**
2.1. Parte a. 3
2.2. Parte b. 6
2.3. Parte c. 8

3. Ejercicio 3 **11**
3.1. Parte a 11
3.2. Parte b 12

1. Ejercicio 1

Se consideran tres redes de interacción proteínas de levadura: Red de copertenencia a complejos proteicos (AP-MS); Red de interacciones binarias (Y2H); y una red obtenida de la literatura (LIT), todas obtenidas del *Yeast Interactome Database*, y mostradas en la figura 1.

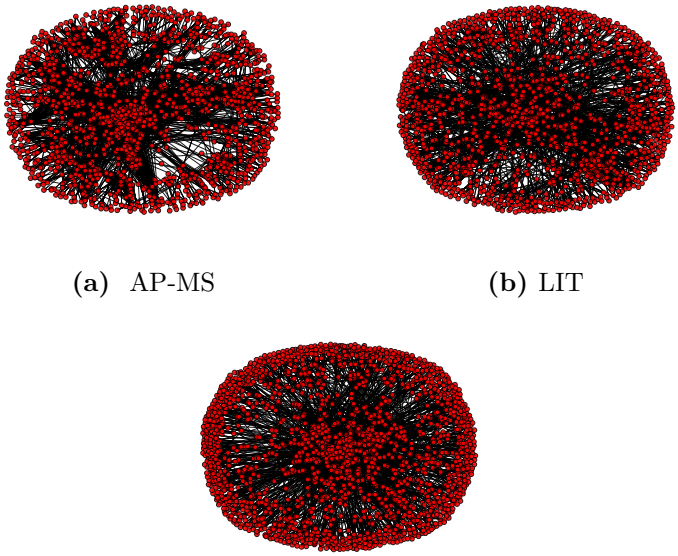
Graficamente se puede observar que la red Y2H tiene un nucleo bien definido y un gran número de nodos conectados a él, esto significa que es posible diferenciar dos grupos proteicos: Mediadoras (interaccionan con un gran número de proteínas) y Mediadas (realizan algún tipo de función con ayuda de las mediadoras). El grupo de las Mediadoras (Nucleo) es probable que esté formado exlusivamente por enzimas.

Por otro lado en grafo AP-MS se pueden observar pequeños nucleos/clusters muy conectados entre si y menos conectados con el resto. Estos nucleos dispersos por la red son los complejos proteicos que busca carecterizar la red.

Que decir de la red LIT ??

Una caracterizacion global y cuantitativa de las redes se muestra en la Tabla I. En ella se muestran diferentes observables globales **Son realmente dirigidas ? analisis simple pero se debe tener claro si se está analizando lo correcto...**

Diagrama de venn ???



(a) AP-MS **(b)** LIT

(c) Y2H

Figura 1: Esquematización de los grafos de cada red estudiada. Se utilizó un *layout* estilo Frunchterman-Reigold para la representación.

Observables	AP-MS	LIT	Y2H
Red dirigida	Si	Si	Si
Tiene loops	No	Si	Si
N^o nodos N	1622	1536	2018
N^o enlaces L	9070	2925	2930
Densidad	0.0034	0.0014	0.0007
Diametro	10	8	11
Grado in k_{in}			
medio $\langle k_{in} \rangle$	5.59	1.90	1.45
maximo máx($\{k_{in}\}$)	111	23	66
minimo mín($\{k_{in}\}$)	0	0	0
Grado in k_{out}			
medio $\langle k_{out} \rangle$	5.59	1.90	1.45
maximo máx($\{k_{out}\}$)	85	35	38
minimo mín($\{k_{out}\}$)	0	0	0
Coeficiente de Clusterización			
medio/local $\langle C \rangle$	0.0741	0.4556	0.0970
triangular/global $C_{\triangle}$	0.6185	0.3461	0.0236

Tabla I: Observables para las tres redes de interacción proteíca de levadura.

2. Ejercicio 2.

La red de este ejercicio trata de una comunidad de delfines de Doubtful Sound, Nueva Zelanda. La comunidad, que se constituye de 62 ejemplares

identificados por una marca en la aleta dorsal, fue fotografiada entre 1995 y 2001. Los datos consisten en un número identificador para cada delfín, su nombre, su género (para 4 ejemplares no está especificado) y la información sobre entre qué pares de ejemplares se forma un link. La red es no dirigida, y contiene un total de 159 links, donde se establece que existe un link entre aquellos individuos que fueron vistos juntos de forma más frecuente que la esperada aleatoriamente, es decir, por un criterio de "compañía preferida".[1]

El ejercicio se dividió en tres partes: en la parte (a), exploramos diferentes layouts para la visualización del grafo; en la parte (b), nos preguntamos si la red es homofílica, es decir, si un delfín tiende a formar enlaces con aquellos ejemplares del mismo género; y en la parte (c), proponemos un método para descomponer la red en dos comunas quitando la menor cantidad de nodos posibles.

2.1. Parte a.

En esta primera parte del ejercicio, exploramos diferentes layouts para visualizar la red delfines. En la figura 2 observamos el resultado de graficar el grafo con el Fruchterman - Reingold layout. El algoritmo para realizar este layout se basa en asignarles fuerzas de interacción ficticias a los nodos. Típicamente se basa en que los nodos ligados tengan una fuerza de atracción análoga a la fuerza de un resorte, sumado a una fuerza de repulsión entre todos los nodos, análoga a la interacción coulombiana entre partículas cargadas idénticamente.[2] Este estilo de layout nos permitió visualizar la existencia de dos comunas de delfines, ligadas a través de unos pocos nodos.

Otros layouts que nos aportan la misma intuición son el DrL layout y el Kamada Kawai (fig. 3), que también están basados en la asignación de fuerzas ficticias. Preferimos el Fruchterman - Reingold layout, ya que los nodos aparecen mejor distribuidos, y permite una mejor visualización de la red. A modo de ejemplo, en la figura 4 incluímos otros layouts: Random, que sitúa los nodos en forma aleatoria, y Multidimensional Scaling, que se basa en una proyección matricial a un espacio de baja dimensionalidad, que no nos aportaron una buena visualización.

[1] D. Lusseau, The emergent properties of a dolphin social network, Proc. R. Soc. London B (suppl.) 270, S186-S188 (2003).

[2] https://en.wikipedia.org/wiki/Graph_drawing

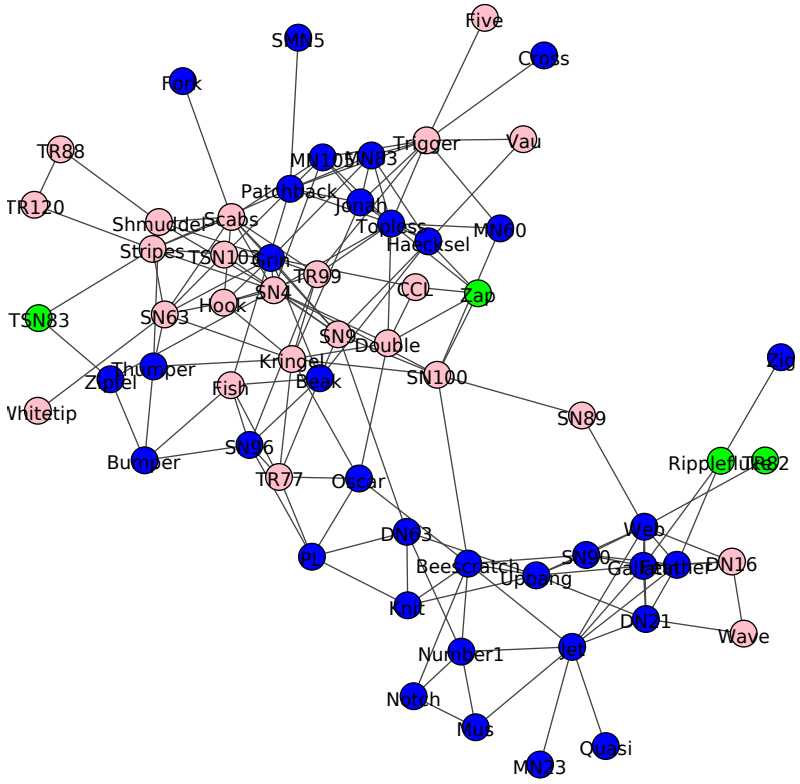


Figura 2: Fruchterman - Reingold layout. Los colores de los nodos se refieren al sexo del delfín: azul, macho; rosa, hembra; verde, sexo no indicado en el dataset. A partir de este layout se puede intuir la existencia de dos comunas de delfines ligadas por pocos nodos.

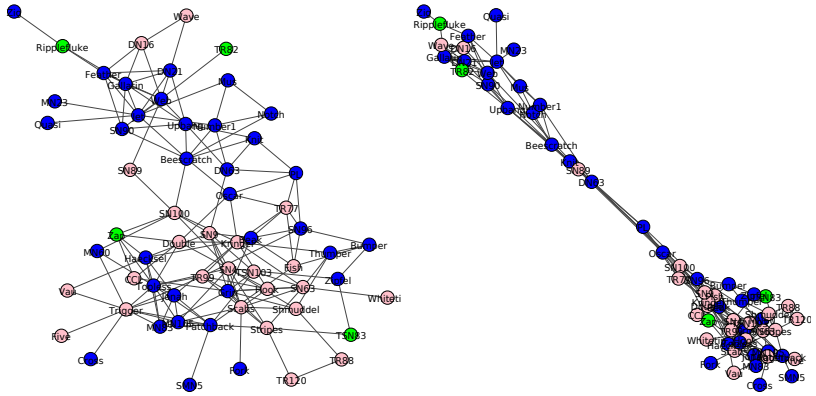


Figura 3: Layouts alternativos: Kamada Kawai izquierda, DrL derecha. Estos también dan la información de una red con dos comunas de tamaño comparable.

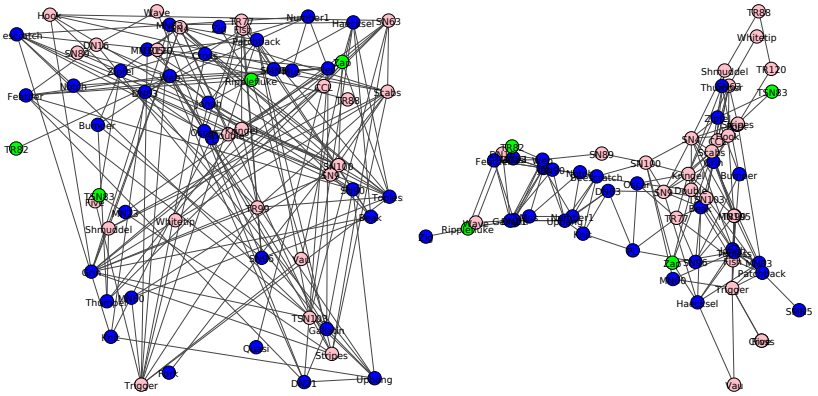


Figura 4: Layouts que no permiten una buena visualización del problema: Random izquierda, Multidimensional Scaling derecha.

2.2. Parte b.

En esta parte nos propusimos estudiar si la red es de carácter homofílica, es decir, si un dado nodo tiende a ligarse con nodos que comparten una característica con él, como es en este caso el género de los delfines. Para ello consideremos la hipótesis nula, es decir, que la asignación de género a un dado nodo es totalmente independiente de la topología de la red, y comparamos con lo presente en el dataset.

La metodología empleada fue la siguiente: sorteamos el género de los delfines manteniendo inalterable la topología de la red y manteniendo constante la cantidad de delfines machos, hembras, y género no específicado de la red original. Generamos 10^6 realizaciones distintas, y para cada caso calculamos la cantidad de links entre delfines de distinto género (sin tomar en cuenta los links entre pares de nodos que incluya un género indefinido). El resultado es la distribución de la figura 5. De dicha distribución obtuvimos un valor medio de links entre géneros $< m >= 68 \pm 7$. Dado que el número de links totales (N') es $N' = 159$, la fracción de links entre géneros es:

$$(\frac{< m >}{N'})_{hip.nula} = 0{,}43 \pm 0{,}04$$

En la figura 5, incluímos la cantidad de links entre géneros de la red real, que en principio se observa mucho menor que la media de la distribución.

Por otro lado podemos estimar analíticamente la fracción de links entre géneros que habría para una red aleatoria. Si tomamos que la probabilidad de escoger un link entre un macho y una hembra es $2\rho_M\rho_F$, donde ρ son las densidades de cada género en la red real (M para macho, F para hembra), entonces la probabilidad de tomar m links de esta característica es:

$$P(m) = \binom{N'}{m}(2\rho_M\rho_F)^m(1 - 2\rho_M\rho_F)^{N'-m} \tag{1}$$

donde $N' = N(N-1)/2$, que es el número total de links que se pueden formar en una red de N nodos. Con esta distribución, el valor medio de links y la desviación standard resultan:

$$< m > = 2N'\rho_M\rho_F$$
$$std(m) = (2N'\rho_M\rho_F(1 - 2\rho_M\rho_F))^{1/2}$$

Por lo tanto, la fracción de links ($< m > /N'$) para las densidades de género actuales, estimado mediante la ecuación anterior, resulta:

$$(\frac{< m >}{N'})estimado = 0{,}43 \pm 0{,}01$$

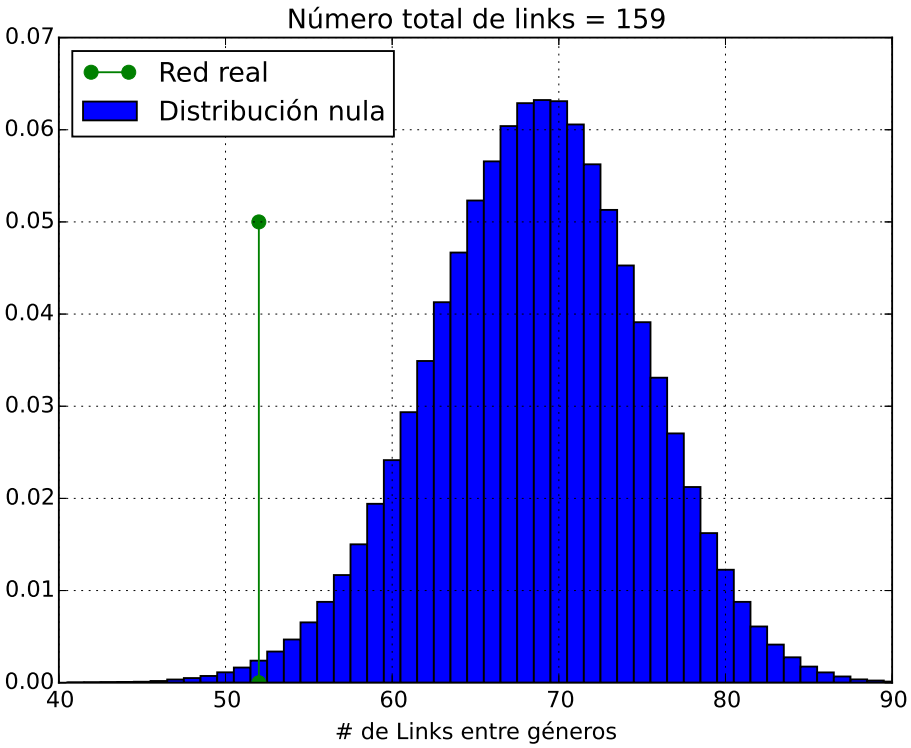


Figura 5: Histograma de links entre géneros sorteando sobre la topología de la red. En verde, cantidad de links entre géneros de la red actual (la altura de la barra no representa nada especial, es solo indicativa).

Observamos que el valor estimado y el obtenido sorteando al azar el género de los delfines coindicen.

Como puede deducirse de la figura 5, la probabilidad de que se obtengan la cantidad de links entre géneros de la red actual dada la hipótesis nula es menor a 0,005, por lo que descartamos la hipótesis nula, y concluimos que la red es homofílica: la cantidad de links entre géneros es mucho menor que el valor esperado si la distribución de géneros fuera aleatoria, lo que implica que hay más cantidad de enlaces entre delfines del mismo sexo que entre ejemplares de distinto género, y además la probabilidad que el valor actual se dé por simple aleatoriedad es muy baja.

2.3. Parte c.

Como último punto, proponemos un método para dividir la red en dos componentes de tamaño comparable. Debido a que suponemos que la red se constituye principalmente de dos comunas ligadas por pocos nodos, consideramos que el cálculo del betweenness no dará información sobre dichos nodos. El betweenness tiene en cuenta la cantidad de caminos cortos que pasan por un dado nodo. Su remoción implicaría un aumento en la distancia entre nodos. Formalmente, el betweenness de un nodo se calcula como:

$$Bet(i) = \sum_{j,k} \frac{b_{jik}}{b_{jk}} \tag{2}$$

donde b_{jk} es el número de caminos cortos que van desde j hasta k, y b_{jik} es el número de caminos cortos que van desde j hasta k, que pasan por i.

Removiendo los nodos con mayor betweenness, basta remover 4 nodos para descomponer la red en dos conjuntos no conexos, como puede observarse en la figura 6. Si lo comparamos con el caso de remover nodos al azar sin más criterio, es muy díficil obtener este resultado. En la figura 7, mostramos el tamaño del componente conectado más grande del sistema a medida que removemos nodos con los dos métodos. Podemos observar que siguiendo el criterio de remover siempre el nodo con mayor betweenness, el tamaño del componente más grande tiene caídas abruptas, lo que indica una partición en el grafo de componentes de tamaño considerable. Sin embargo, al remover al azar, se observa que el componente más grande va disminuyendo su tamaño de forma suave con la remoción de cada nodo particular.

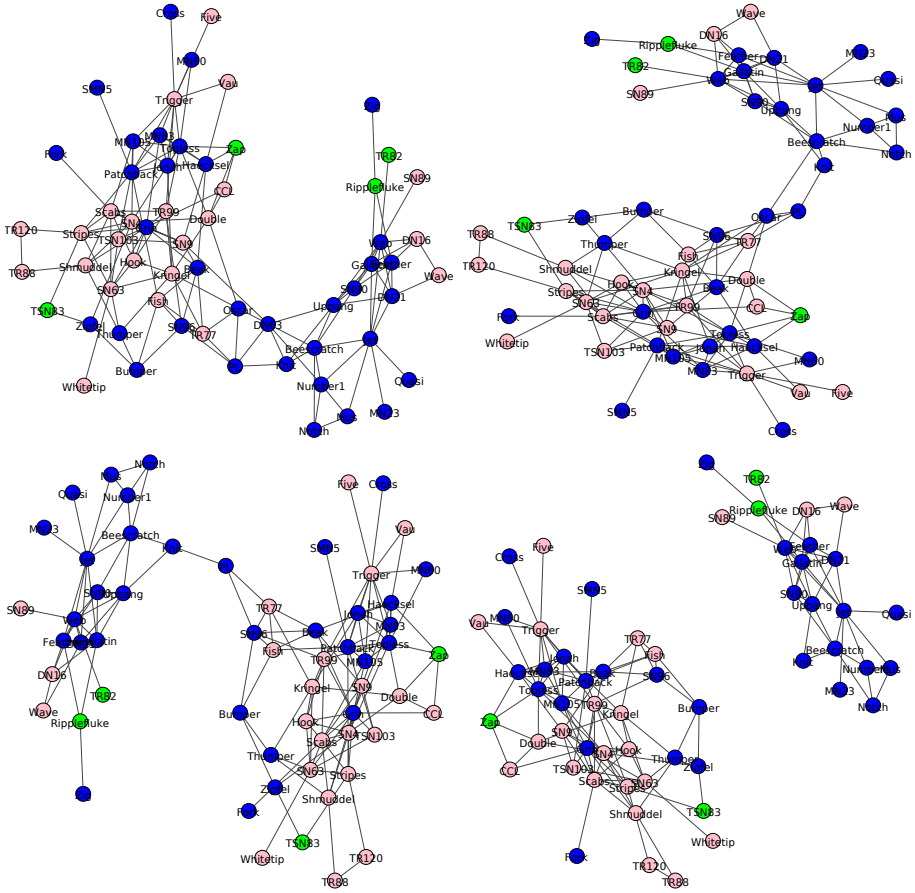


Figura 6: Layout de la red al remover el nodo con mayor betweenness: remoción de izquierda a derecha, y de arriba hacia abajo. Con solo remover 4 nodos, la red se descompone en dos subgrafos de tamaño comparable.

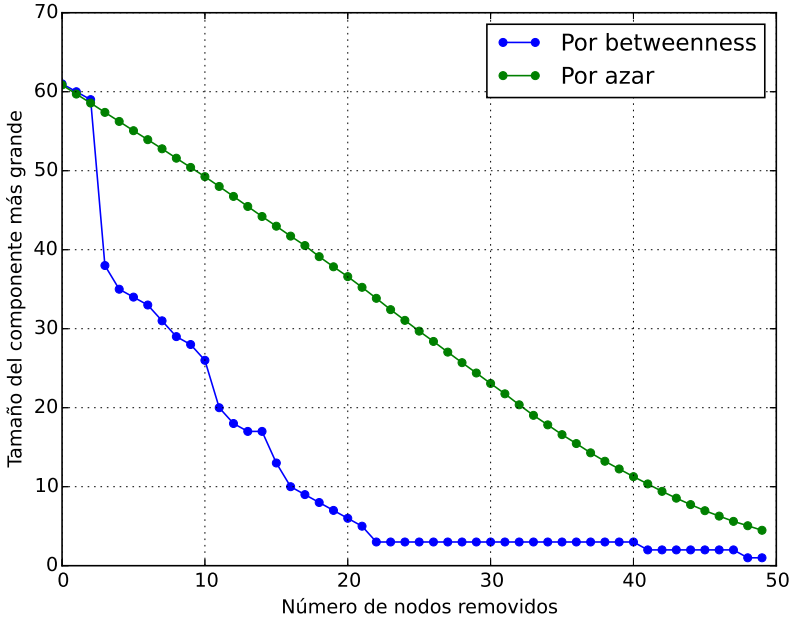


Figura 7: Tamaño del componente más grande a medida que se remueven nodos según betweenness, o al azar, respectivamente. Observar la caída abrupta al remover 4 nodos de mayor betweenness. La curva correspondiente a la forma azarosa es un promedio de 1000 configuraciones.

3. Ejercicio 3

En esta sección consdideramos una red (ver archivo `as-22july06.gml`) creada por Mark Newman que contiene la estructura de los sistemas autónomos de internet relevada a mediados del 2006.

3.1. Parte a

Primero hallamos todos los grados k de la red, usando la rutina `igraph.Graph.Read_GML()` de la libreria `igraph` (version 0.7.1) de Python. La distribución P_k de dichos valores de grado se observa en la figura 8, donde claramente observamos un fuerte comportamiento tipo power-law, a lo largo de trés décadas en k. Exploramos visualmente la tendencia de P_k en otras escalas lineales y semi-logarítmicas, pero no fueron las más apropiadas para caracterizarla.

3.2. Parte b

A continuación, repetimos el cálculo del ajuste para el exponente de la distribución power-law $P_k \sim k^{-\alpha}$, pero esta vez usando la rutina `igraph.ipower_law_fit()`, el cual entrega valores de p-value=0,5, k_{min}=3,9, y del exponente $\alpha = 1{,}37$.

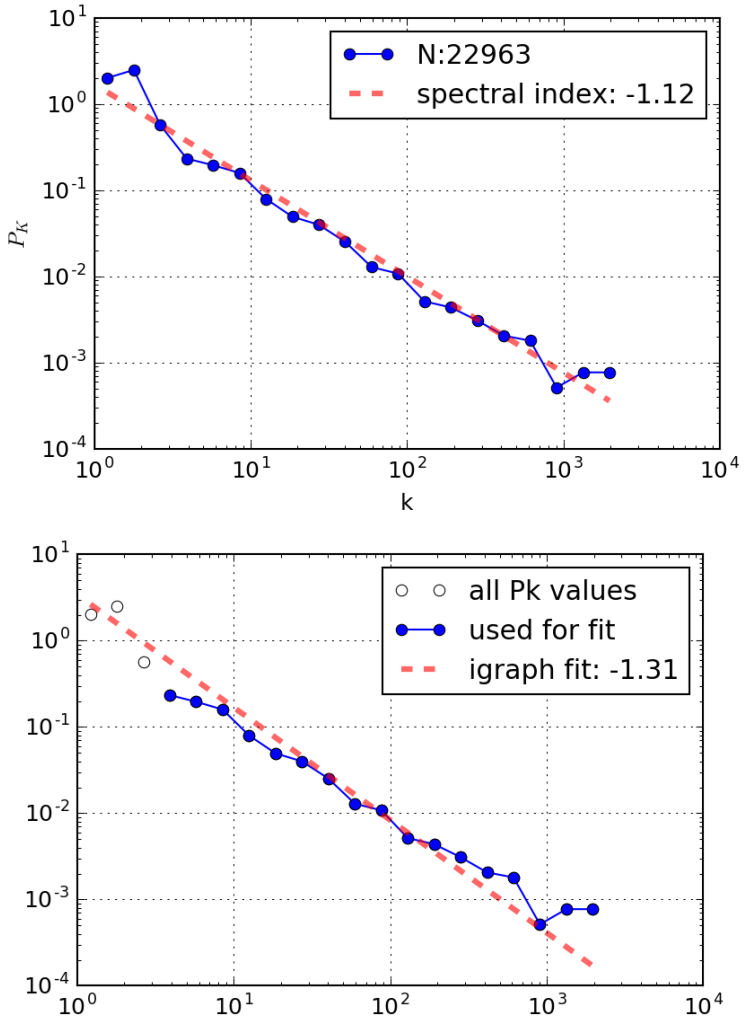


Figura 8: Distribución de probabilidad P_k de los valores de grado k de la red de Mark Newman. Observamos que la distribución tiene fuerte comportamiento tipo power-law, especialmente en el rango $k \sim (3 - 400)$.