Trabajo Práctico 1: Conceptos básicos

Emanuel Ferreyra, Bruno Kaufman, Ariel Salgado

26 de septiembre de 2018

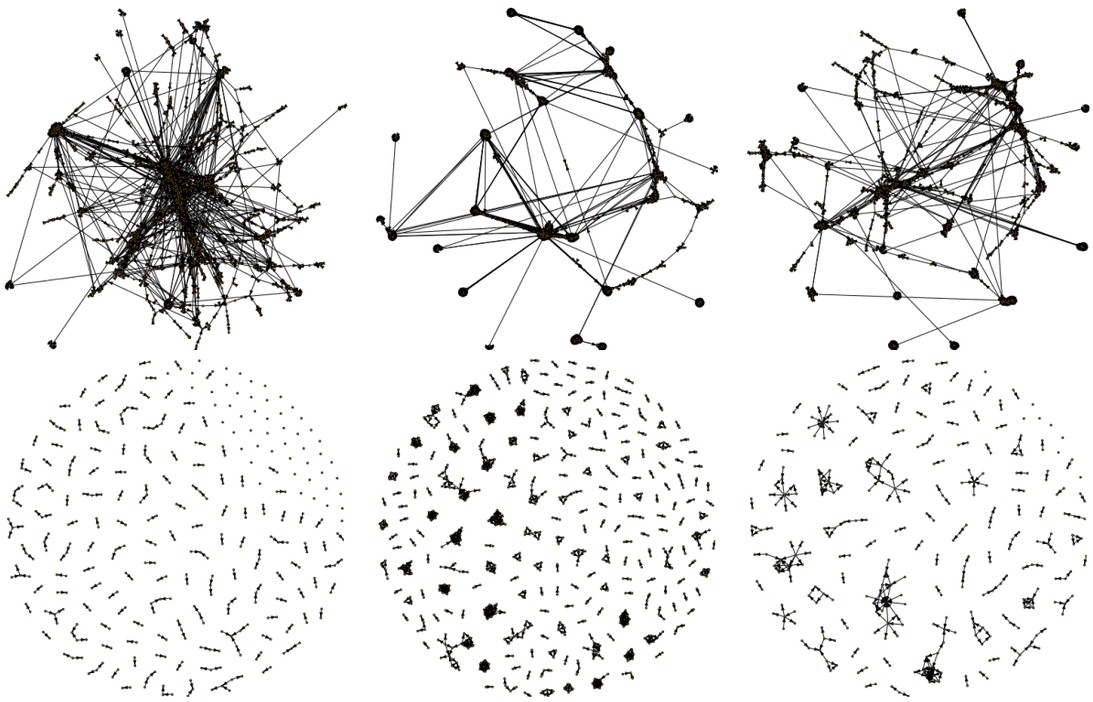
1. Problema 1

En este punto se pide analizar tres redes complejas de interacción de proteinas. La primera se trata de interacciones binarias entre proteinas, la segunda de co-pertenencia a complejos, y la tercera una obtenida del proyecto Yeast Interactome Database (YID).

1.1. Punto 1.a.

En este inciso se solicita visualizar las redes de forma informativa acerca de su estructura. Las tres redes se visualizan en la tabla 1.

Tabla 1: **Superior:** Componente gigante perteneciente a la red del proyecto YID (Y2H), de interacciones binarias (APMS), y de co-pertenencia (LIT), de izquierda a derecha. **Inferior:** misma información respecto a las componentes pequeñas de las redes.



Se pueden observar algunas cosas en base a estos diagramas. En primer lugar, observando sólamente a los componentes pequeños podemos ver que en la red Y2H, no presentan clustering considerable. En contraste, la red APMS presenta clustering significativo, y la red LIT también, aunque no tanto como la red APMS.

Por otro lado, observando el componente gigante se puede ver que la mayoría de los links de la red Y2H tienden hacia un mismo lugar en la red, mientras que la red APMS se presenta menos centralizada y la LIT toma un punto medio entre las dos.

1.2. Punto 1.b.

En este inciso se pide resumir ciertas características de red en una tabla. La siguiente se puede ver en la tabla 2.

Tabla 2: Cantidades de red asociadas a las redes correspondientes.

Red	**Y2H**	**APMS**	**LIT**
Número de nodos	2018	1622	1536
Número de ejes	2930	9070	2925
Grado medio	2.904	11.184	3.809
Densidad	$1.440\ 10^{-3}$	$6.899\ 10^{-3}$	$2.481\ 10^{-3}$
Clustering global	$2.361\ 10^{-2}$	$6.186\ 10^{-1}$	$3.462\ 10^{-1}$
Clustering local medio	$9.700\ 10^{-2}$	$7.410\ 10^{-1}$	$4.556\ 10^{-1}$
Diámetro	14	15	19

Se puede ver que efectivamente la red Y2H presenta un clustring significativamente menos a las otras dos, y que la APMS tiene uno mayor a la red LIT, aunque están en el mismo órden de magnitud.

Además, es interesante ver que el diámetro nos informa que para llegar de un extremo al otro de la red LIT, el número máximo es apreciablemente mayor que en las otras dos. Esto no es fácilmente visible en los diagramas vistos en la tabla 1

1.3. Punto 1.c.

Las interacciones reportadas son todas de naturaleza biomolecular. Por un lado, es razonable que exista una sola componente gigante para estas redes, ya que las redes biológicas presentan una distribución de grado libre de escala, consistente con un mecanismo de attachment preferencial: en otras palabras, cuanto mayor el grado de un nodo, mayor la probabilidad de que tenga un dado enlace con otro. Entonces, la componente más grande siempre atraerá la mayoría de los enlaces, y se volverá más grande, llegando a ser la única componente gigante. Además, se espera en todos los casos la presencia de pocos hubs con grado significativamente mayor a la mayoría, al ser la distribución de grado libre de escala.

Por otro lado, es razonable que una red de co-pertenencia (LIT) tenga una clusterización significativa, ya que muchas proteínas pueden pertenecer a un mismo complejo, y así todas estarán conectadas entre si formando un clique. Suponiendo que esta co-pertenencia influye sobre las interacciones binarias, tiene también sentido que una clusterización alta se presente en la red de interacciones binarias (APMS).

Además, la red de copertenencia (LIT) presenta un diámetro más grande que las otras dos. Esto es razonable, porque cuando se trata de interacciones binarias (APMS, Y2H), es posible que interactúen dos proteínas que no estén en el mismo complejo, agregando conexiones y disminuyendo el diámetro.

Esto también explica el grado medio mayor de la red de interacciones binarias (APMS), pero no la del interactoma de la levadura (Y2H), que presenta el grado medio más pequeño de todas. Esto puede ser por falsos negativos: al observarse sólamente aquello que ocurre en el núcleo, van a haber muchas interacciones que se pierden. Esto también explicaría el clustering relativamente pequeño del interactoma de la levadura.

2. Problema 2

En el segundo problema se propone analizar una red social de 62 delfines de Nueva Zelanda. La red consta de 34 machos, 24 hembras y 4 delfines de sexo desconocido. La base de datos establece 159 vinculos entre los delfines de la red.

2.1. Punto 2.a.

El primer objetivo es comparar diferentes opciones de layout para graficar la red, comparando ventajas y desventajas. Para esto empleamos el paquete `igraph` de `R`. El mismo provee distintas opciones de layout a partir de funciones predeterminadas. Presentamos para ejemplificar tres layouts: `layout-with-fr`, `layout-as-tree` y `layout-on-grid`. El primero construye el layout a partir de un cálculo de equilibrio de fuerzas, el segundo intenta representar al grafo como un árbol, permitiendo que se le indique un nodo semilla, y el tercero posiciona los nodos sobre una grilla regular. En las figuras 1, 2, y 3 se pueden observar los mismos. Los nodos rojos representan los machos, los azules las hembras, y los negros los de sexo desconocido. Ninguno de los últimos dos facilita la visualización del gráfico, indicando que las estructuras propuestas no son representativas de la estructura del grafo. Por otro lado, `layout_with_fr`, al tener directamente en cuenta los vinculos de la red para su construcción, nos permite observar que hay dos grupos de delfines bastante diferenciados, uno en el que predominan las hembras (y dominan los colores azules) y otro donde predominan los machos (mayoría de colores rojos).

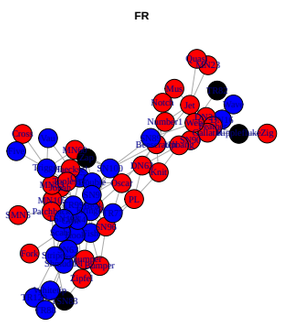


Figura 1: Red de los delfines, graficada según `layout_with_fr`.

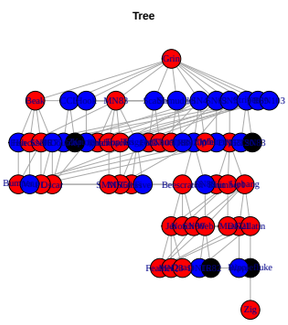


Figura 2: Red de los delfines, graficada según `layout_as_tree`.

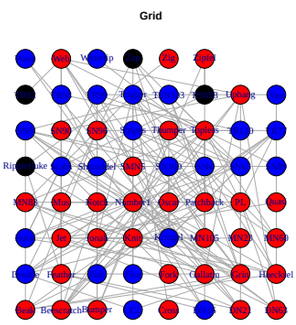


Figura 3: Red de los delfines, graficada según `layout_on_grid`.

2.2. Punto 2.b.

En este punto se propone testear la hipótesis que sugiere `layout_with_fr`: que hay más vinculos entre delfines del mismo sexo (vinculos homofilos) que entre delfines de distinto sexo (vinculos heterófilos). Para esto, calculamos la fracción de ejes que conectan delfines de distinto sexo como

$$f_{mf} = \frac{1}{M} \sum_{ij} A_{ij} \delta_{s_i,m} \delta_{s_j,f} \tag{1}$$

donde s_i es el sexo del delfin i y A_{ij} es la matriz de adyacencia del red, y $M = \sum_{ij} A_{ij}/2$ es la cantidad de ejes en la red. De forma análoga, calculamos las fracciones de ejes que conectan hembras con hembras y machos con machos:

$$f_m = \frac{1}{2M} \sum_{ij} A_{ij} \delta_{s_i,m} \delta_{s_j,m} \tag{2}$$

$$f_f = \frac{1}{2M} \sum_{ij} A_{ij} \delta_{s_i,f} \delta_{s_j,f} \tag{3}$$

En todos los casos, consideramos únicamente los defines con sexo conocido. Obtenemos los valores de $f_{mf} = 0{,}327$, $f_m = 0{,}377$ y $f_f = 0{,}226$, quedando aproximadamente un 7 % de los ejes sin considerar por no conocerse el sexo. Para saber si estos valores son altos o bajos, el ejercicio propone reasignar de forma azarosa los sexos de cada delfín, y calcular las tres fracciones para cada asignación de sexo. Realizando 1000 reasignaciones, obtenemos tres distribuciones de valores. Podemos observar estos en las figuras 4, 5, y 6. Obtenemos valores esperados de $\widehat{f_{mf}} = 0{,}43 \pm 0{,}04$, $\widehat{f_m} = 0{,}30 \pm 0{,}04$ y $\widehat{f_f} = 0{,}15 \pm 0{,}03$, de

donde vemos que los valores observados se encuentran entre 1.8 y 2.8 sigmas de distancia de los esperados. Si calculamos los p valores en cada caso, obtenemos $p_{mf} = 0{,}005$,$p_m = 0{,}029$ y $p_f = 0{,}015$, indicando una muy baja probabilidad de que por azar la estructura observada se forme.

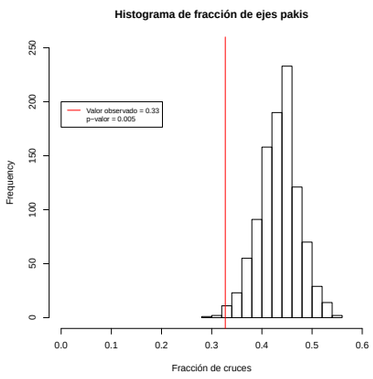


Figura 4: Histograma de fracción de ejes que conectan delfines de sexos distintos.

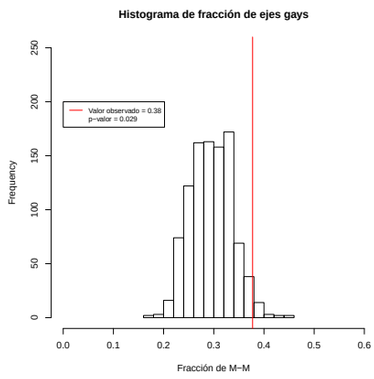


Figura 5: Histograma de fracción de ejes que conectan defines macho con macho.

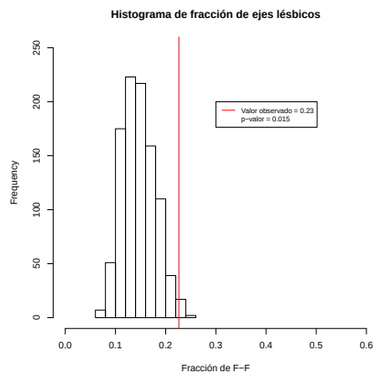


Figura 6: Histograma de fracción de ejes que conectan defines hembra con hembra.

2.3. Punto 2.c.

Por último, el ejercicio propone comparar distintas estrategias para remover nodos de la red, basandose en la estructura de la misma. Para esto proponemos tres estrategias: extraer nodos según su grado, su coeficiente local de clustering, y su betwenness (una medida de la cantidad de caminos mínimos que atraviezan un eje). En la figura 7 podemos observar el tamaño de la primer y segunda componente. Elegimos la estrategia más apropiada viendo cual asemeja los tamaños más prontamente. Vemos que la estrategia más exitosa es la que extrae por betweenness. Seguida de ella está extraer por grado, que tiene éxito cuando las componentes ya son más pequeñas. Por último, extraer según el coeficiente de clustering rinde resultados de tamaño mucho menor, habiendo desperdiciado la mayoría de las eliminaciones en formar componentes pequeñas en vez de partir la componente gigante en dos similares.

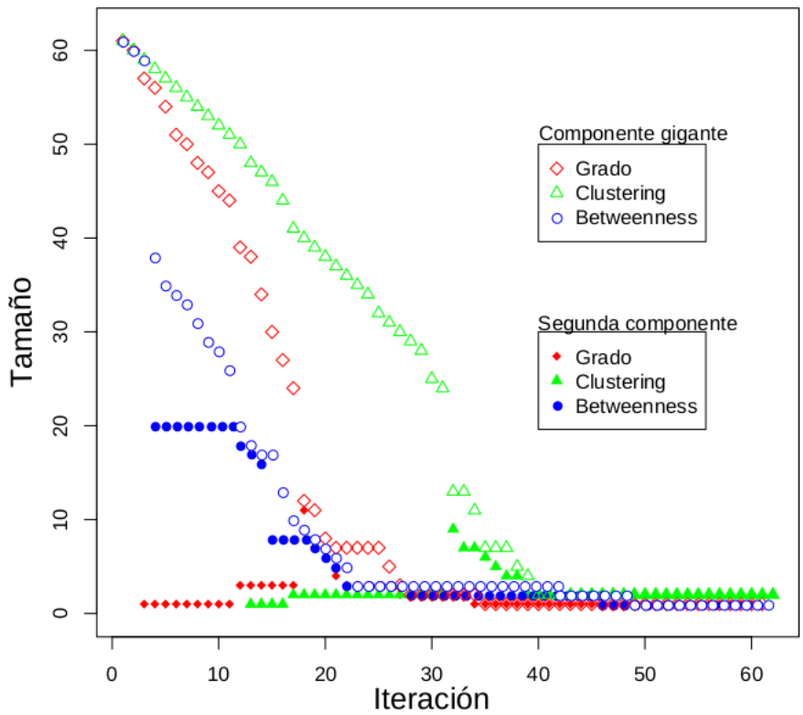


Figura 7: Decaimiento en el tamaño de la componente gigante y la segunda más grande a medida que se aplican algoritmos de destrucción basados en la eliminación de nodos por el grado de un nodo, por su coeficiente de clustering local, y por su betweenness. Cuando los puntos huecos de un color equivalen a los puntos llenos, se considera que se tienen dos componentes de igual tamaño como gigantes.

3. Problema 3

En este ejercicio estudiamos la distribución de grados P_k como función de k. Nuevamente utilizamos las funciones del paquete `igraph` de `R`, consiguiendo primero los grados de cada nodo y luego utilizando el comando `hist` que computa un histograma a partir del vector de grados.

En la siguiente tabla están los gráficos con distintas alternativas, con bineado lineal o logarítmico y utilizando diferentes escalas.

En las figuras 4 a 6 podemos ver la distribución de grados utilizando un bineado lineal. Se se aprecia la forma de la distribución aunque es complicado cuantificar si un ajuste libre de escala será coherente. En figuras 12 y 13 se utilizó un bineado logarítmico en base 2, y en la escala logarítmica se aprecia bien el carácter libre de escala de dicha distribución.

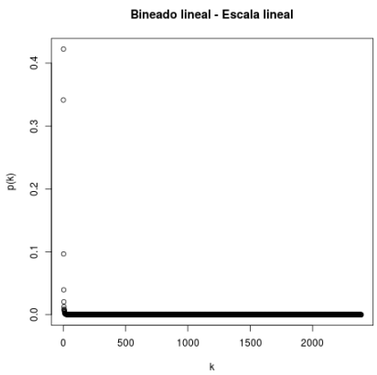


Figura 8: Bineado Lineal - Escala lineal

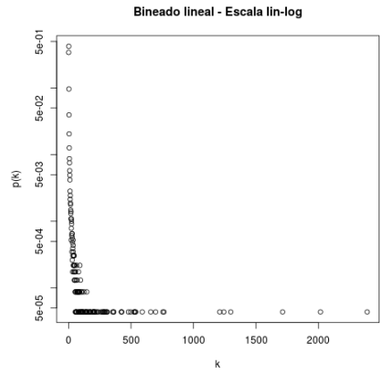


Figura 9: Bineado Lineal - Escala lineal en x logarítmica en y

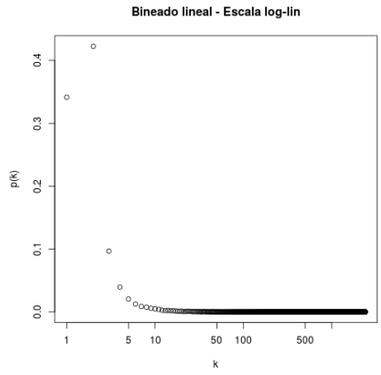


Figura 10: Bineado Lineal - Escala logarítmica en x lineal en y

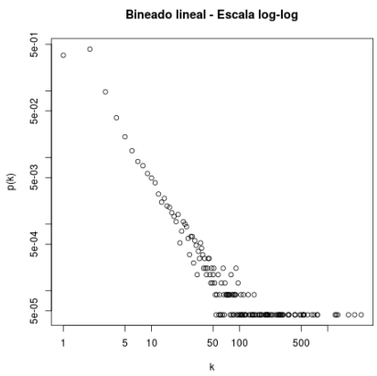


Figura 11: Bineado Lineal - Escala logarítmica

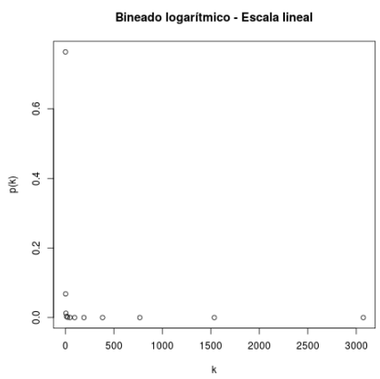


Figura 12: Bineado logarítmico - Escala lineal

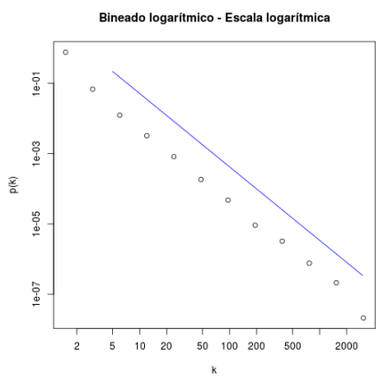


Figura 13: Bineado logarítmico - Escala logarítmica - Con recta del ajuste PL

En la figura 13 se agregó la recta de ajuste libre de escala obtenida vía el comando *fit_power_law* utilizando una malla de 3000 puntos a partir del k_{min} que en nuestro caso toma valor 5. El valor estimado es $\alpha = 2{,}097157$ ajustandose correctamente aunque corrido hacia arriba. Esto se debe a que el parámetro de normalización, es decir, la ordenada al origen, se calcula teniendo en cuenta todos los valores de grado, no solo los mayores al k_{min}.

4. Problema 4

En el cuarto problema se propone realizar distintas medidas para analizar la asortatividad de una red. La asortatividad representa la tendencia de un nodo a parecerse topológicamente a sus vecinos (por ejemplo en términos de grado). Para esto se toman en consideración dos redes: *netscience* y *as_july_22*. La primera es una red de colaboraciones científicas y la segunda de internet.

4.1. Punto 4.a.

El primer cálculo que se propone para observar la asortatividad es calcular el valor de $k_{nn}(k)$, que es el valor medio del grado de los vecinos de los nodos de grado k. En las figuras 14 y 15 podemos observar los valores de $k_{nn}(k)$ en función de k. Siguiendo la propuesta de Newmann, podemos hacer un ajuste lineal de $\log(k_{nn}(k) \sim \log(k)$, para obtener el coeficiente μ, donde

$$k_{nn}(k) = Ak^{\mu} \tag{4}$$

Siguiendo este modelo, obtenemos los valores

$$\mu_{july} = -0{,}44 \pm 0{,}04 \tag{5}$$

$$\mu_{netsci} = 0{,}31 \pm 0{,}07 \tag{6}$$

de donde podemos ver que la red *as_july_22* es desasortativa (es decir, que nodos con grados altos tienden a juntarse con nodos con grado bajo), y la red *netscience* es asortativa, juntandose nodos con grados altos con otros con grados altos. Este resultado pareciera tener sentido bajo la siguiente visión: en la red *as_july_22* tenemos nodos con grado muy alto, que son centrales a la conectividad de la red de internet. Todos los nodos de grado bajo se conectan a ellos ya que funcionan como una central de comunicación. En cambio, en la red *netscience*, que es de coautoria de papers, un nodo con grado alto representa un científico muy productivo. En esta red es esperable que dos científicos trabajen juntos mientras más productivos sean, por lo cual tiene una asortatividad alta.

A continuación, se propone medir la asortatividad con otro coeficiente, propuesto por Newman, según la ecuación:

$$r = \frac{\sum_{ij}(A_{ij} - k_i k_j/2m)k_i k_j}{\sum_{ij}(k_i\delta_{ij} - k_i k_j/2m)k_i k_j} \tag{7}$$

Este coeficiente permite medir la tendencia lineal, pero en escala natural (no logaritmica). Para ambas redes, obtenemos en este caso:

$$r_{july} = -0{,}19 \tag{8}$$

$$r_{netsci} = 0{,}46 \tag{9}$$

de forma tal que ambos coeficientes apuntan a las mismas conclusiones.

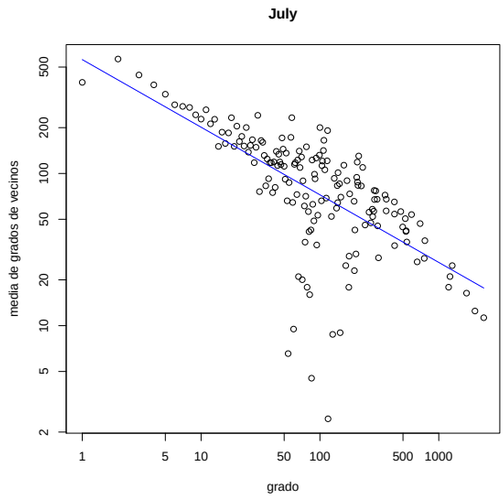


Figura 14: Valores de $k_{nn}(k)$ en función del grado k para la red *as_july_22*. En azul la regresión lineal.

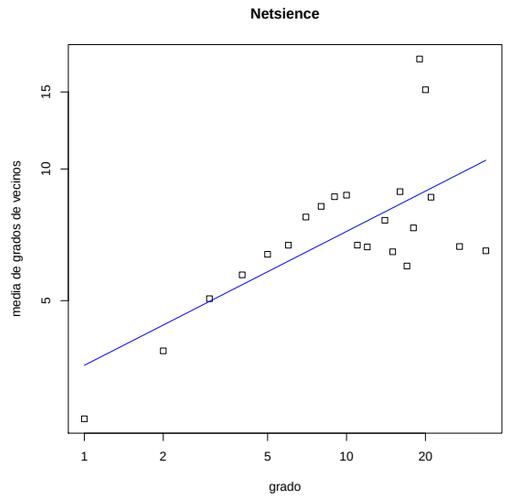


Figura 15: Valores de $k_{nn}(k)$ en función del grado k para la red *netscience*. En azul la regresión lineal.

4.2. Punto 4.b.

Por último, se propone repetir las medidas para las redes de interacción de proteinas de levadura Y2H y AP-MS presentadas en el punto 1. En las figuras 16 y 17 se encuentran graficados los valores de $k_nn(k)$ en función de k con sus regresiones lineales en escala log-log.

En este caso los datos están más dispersos, por lo que la calidad de las regresiones disminuye. Obtenemos los valores:

$$\mu_{Y2H} = -0{,}26 \pm 0{,}15 \tag{10}$$

$$\mu_{APMS} = 0{,}39 \pm 0{,}12 \tag{11}$$

$$r_{Y2H} = -0{,}04 \tag{12}$$

$$r_{APMS} = 0{,}60 \tag{13}$$

donde nuevamente vemos acuerdo entre los coeficientes r y μ, aunque en el caso de la red Y2H la correlación es muy baja. En este caso,vemos que la red Y2H es desasortativa, y la red APMS es asortativa. Dado que la red Y2H se construye a partir de interacciones binarias, únicamente aquellas proteinas que interactuan con muchas otras tendrán muchas conexiones. En cambio, en la red APMS se establecen conexiones entre los grupos de conexiones que pertenecen a un complejo proteico común. Esto facilita la vinculación entre nodos de grado alto en la red APMS, ya que proteinas que interactuan con muchas otras (menos interactuantes), terminaran conectadas igualmente, al encontrarse con alta probabilidad ambas en un complejo proteico grande (ya que ambas son muy interactuantes). Cabe notar que en el caso de APMS es muy marcado como a grados altos hay mucha dispersión en el grado medio de los vecinos, indicando que no todas las *proteinas famosas* se conectan principalmente con otras *proteinas famosas*

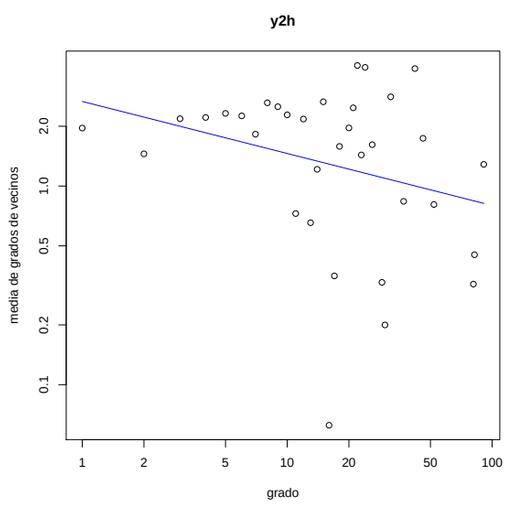


Figura 16: Valores de $k_{nn}(k)$ en función del grado k para la red *Y2H*. En azul la regresión lineal.

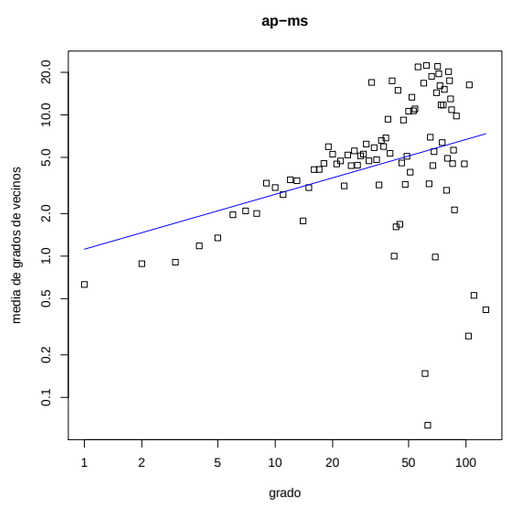


Figura 17: Valores de $k_{nn}(k)$ en función del grado k para la red *AP-MS*. En azul la regresión lineal.