

UNIVERSIDAD DE BUENOS AIRES

INFORME 1

Redes Complejas

*Raúl Barriga
Mariela Celis
Jimmy Masías
Sebastián Pinto*

15 de septiembre de 2016

Índice

1. Ejercicio 1	2
1.1. Caracterización de las redes	2
1.2. Coherencia entre las redes	3
2. Ejercicio 2.	6
2.1. Parte a.	6
2.2. Parte b.	9
2.3. Parte c.	11
3. Ejercicio 3	14
3.1. Parte a	14
3.2. Parte b	14
4. Ejercicio 4	16
4.1. Consideraciones generales	16
4.2. Grado medio del vecindario k_{nn} (revising Barabási [2016]) . .	17
4.3. Coeficiente de Correlación de Grado	18
4.4. Discusión	20

1. Ejercicio 1

Se consideran tres redes de interacción proteínas de levadura *Saccharomyces cerevisiae*: La red de complejos proteicos (red AP-MS) relaciona las proteínas que, formando un complejo proteico (estructura cuaternaria), generan una estructura medible a través de *Mass Spectrometry* (esta técnica es conocida como *Affinity Purification/Mass Spectrometry* [Collins et al., 2007]; La red de interacciones Binarias obtenida a través de la técnica *yeast 2 hybrid* (Y2H) [Yu et al., 2008]; y la red de interacciones reportadas en la literatura [Reguly et al., 2006]. Todas estas redes están disponibles en *Yeast Interactome Database*¹, y son mostradas en la figura 1.

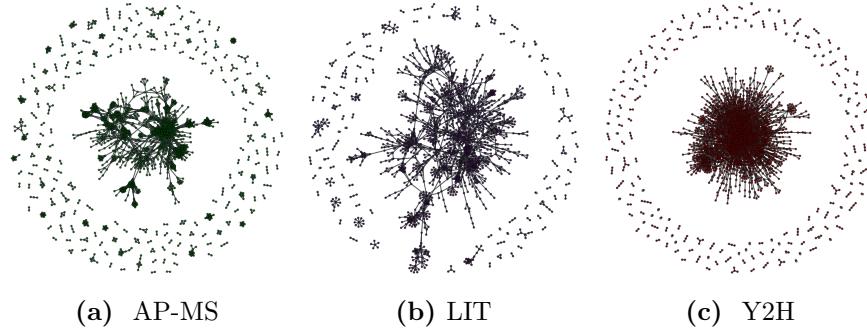


Figura 1: Esquematización de los grafos de cada red estudiada. Se utilizó un *layout* SFDP, de la librería `graph-tool`, y pueden ser reproducidas usando `plot.py <archivo> -f <formato>`

1.1. Caracterización de las redes

Una caracterización global y cuantitativa de las redes se muestra en la Tabla I.

No es de extrañar que, a pesar de que estas 3 redes de proteínas pertenecen al mismo organismo (*Saccharomyces cerevisiae*), estas son distintas topológicamente debido a la manera en que son armadas. La red AP-MS muestra la formación de varios grupos/clusters muy densos que corresponden a los complejos proteicos purificados, esto se ve reflejado en los observables de la tabla I en que, salvo el diámetro, presenta mayores valores que el resto de las redes. Cabe observar que la red de literatura LIT presenta el mayor diámetro, sin embargo es la con menor grado máximo (es decir, es

¹[interactome.dfci.harvard.edu/S_cerevisiae/](http:// interactome.dfci.harvard.edu/S_cerevisiae/)

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

Observables	AP-MS	LIT	Y2H
Nº nodos N	1622	1536	2018
Nº enlaces L	9070	2925	2930
Densidad	0.0068	0.0024	0.0014
Diametro	15	19	14
Grado k			
medio $\langle k \rangle$	11.18	3.80	2.93
maximo máx($\{k\}$)	127	40	91
minimo mín($\{k\}$)	1	1	1
Coefficiente de Clusterización			
medio/local $\langle C \rangle$	0.0710	0.4556	0.0970
triangular/global C_Δ	0.6185	0.3461	0.0236

capaz de relacionar más relaciones entre clusters), debido a la compilación de un gran número de trabajos, pero tiene menor detalle de las interacciones proteína-proteína, debido a que en general los trabajos consultados son específicos y por lo tanto sesgados. En particular al observar la clusterización es importante diferenciar la clusterización media $\langle C \rangle$ y la clusterización triangular C_Δ . La primera caracteriza la media de la clusterización local de cada nodo, es por ello que la red LIT presenta el mayor valor, debido a que, como mencionamos antes, compila trabajos detallados que dan cuenta de pequeños clusters. Tanto la red AP-MS y, aún más, la red Y2H diluyen su clusterización debido a un gran número de *pequeñas interacciones aisladas*. Por otro lado el coeficiente C_Δ es una medida más global, ya que no pesa interacciones binarias puras. En este último caso AP-MS presenta la mayor clusterización, mientras que Y2H la menor.

1.2. Coherencia entre las redes

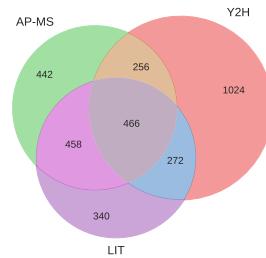


Figura 2: Diagrama de Venn para cobertura entre las tres redes.

En segundo lugar se analizó la cobertura y coherencia entre las interacciones reportadas en las redes. Para ello se realizaron los diagramas de Venn mostrados en la figura 3 (estos pueden reproducirse con el script `venn.py`). Para analizar la cobertura comparamos la intersección de las proteínas reportadas en cada caso. En la figura 2 se muestra la cobertura de cada red y la cantidad de proteinas reportadas por más de una red (intersecciones). Es interesante notar que AP-MS y LIT presentan $\approx 60\%$ de cobertura entre ellas y solo un $\approx 27\%$ y un $\approx 22\%$ de las protínas reportadas, respectivamente, son específicas de cada red. Esto se ve contrastado con el $\approx 50\%$ de especificidad en la red Y2H entre las interacciones.

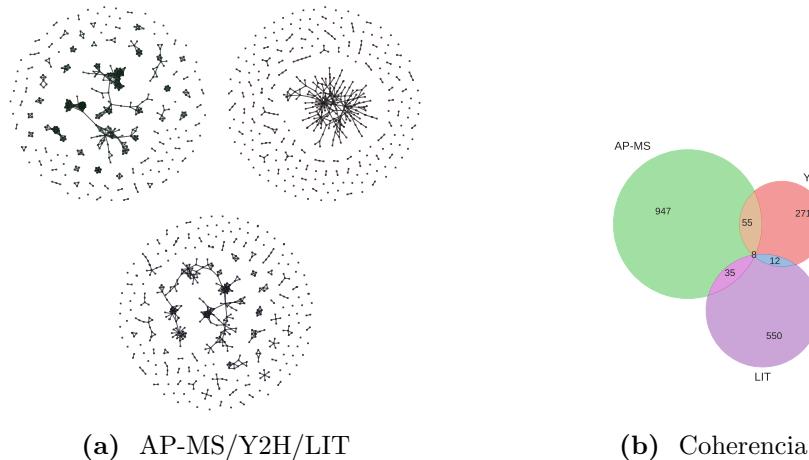


Figura 3: Subgrafos inducidos por cobertura de las 3 redes y coherencia de links entre ellas.

Por otra parte, a partir de la intersección de las proteínas reportadas de las tres redes, se analiza la coherencia de los enlaces entre las proteínas comunes. De la figura 3b. Aquí se puede observar al alta especificidad de cada red respecto al resto (solo 8 links son compartidos por las tres redes).

Adicionalmente se muestran las coherencias a pares y los subgrafos inducidos en la figura 4. Cabe notar que, a pesar que en los subgrafos de la

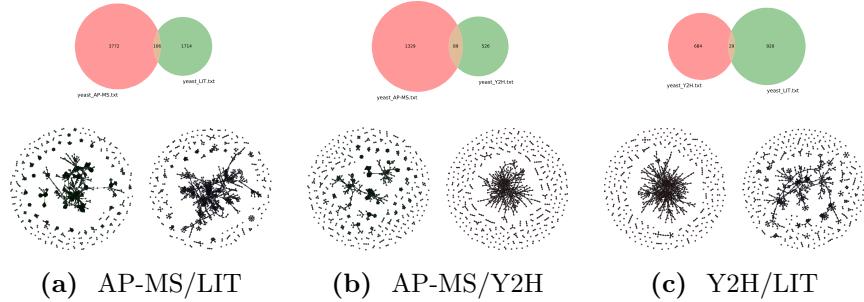


Figura 4: Coherencia entre links para cada par de redes y los respectivos subgrafos de cada red

cobertura entre AP-MS y LIT podrían parecer *similares*, las coherencias entre links en todos los casos son escasas respecto a la cantidad total de links en cada subgrafo.

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”.²

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 5 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.³ 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. 6), 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 7 incluimos 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.

²D. Lusseau, The emergent properties of a dolphin social network, Proc. R. Soc. London B (suppl.) 270, S186-S188 (2003).

³https://en.wikipedia.org/wiki/Graph_drawing

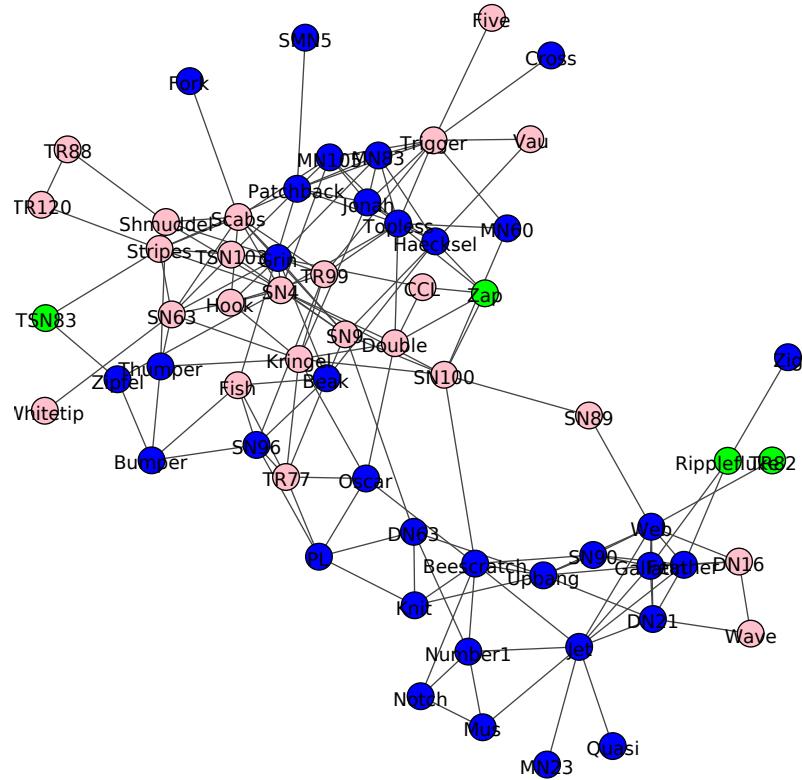


Figura 5: Fruchterman - Reingold layout. Los colores de los nodos se refieren al sexo del delfín: azul, macho; rosa, hembra; verde, sexo no indicado en el dataset. A partir de este layout se puede intuir la existencia de dos comunas de delfines ligadas por pocos nodos.

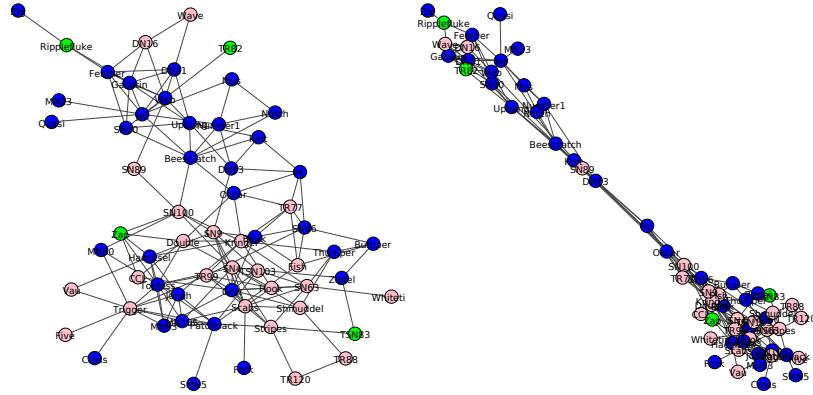


Figura 6: Layouts alternativos: Kamada Kawai izquierda, DrL derecha. Estos también dan la información de una red con dos comunas de tamaño comparable.

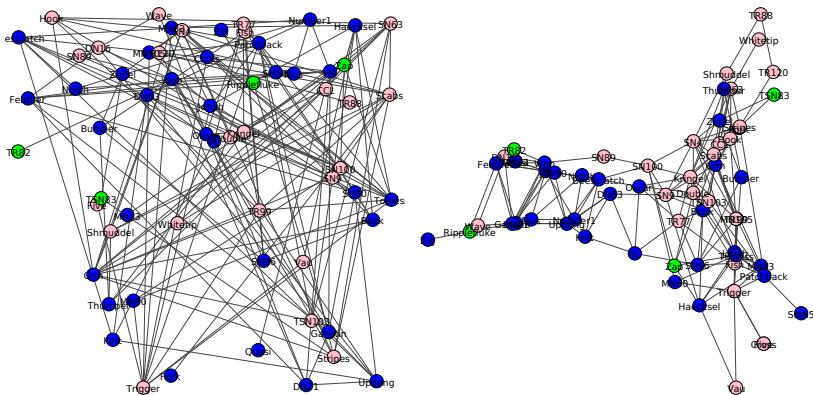


Figura 7: Layouts que no permiten una buena visualización del problema: Random izquierda, Multidimensional Scaling derecha.

2.2. Parte b.

En esta parte nos propusimos estudiar si la red es de carácter homófílico, 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 especificado 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 8. De dicha distribución obtuvimos un valor medio de links entre géneros $\langle m \rangle = 68 \pm 7$. Dado que el número de links totales (N') es $N' = 159$, la fracción de links entre géneros es:

$$\left(\frac{\langle m \rangle}{N'} \right)_{\text{hip.nula}} = 0,43 \pm 0,04$$

En la figura 8, incluimos 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:

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

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:

$$\begin{aligned} \langle m \rangle &= 2N'\rho_M\rho_F \\ std(m) &= (2N'\rho_M\rho_F(1 - 2\rho_M\rho_F))^{1/2} \end{aligned}$$

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

$$\left(\frac{\langle m \rangle}{N'} \right)_{\text{estimado}} = 0,43 \pm 0,01$$

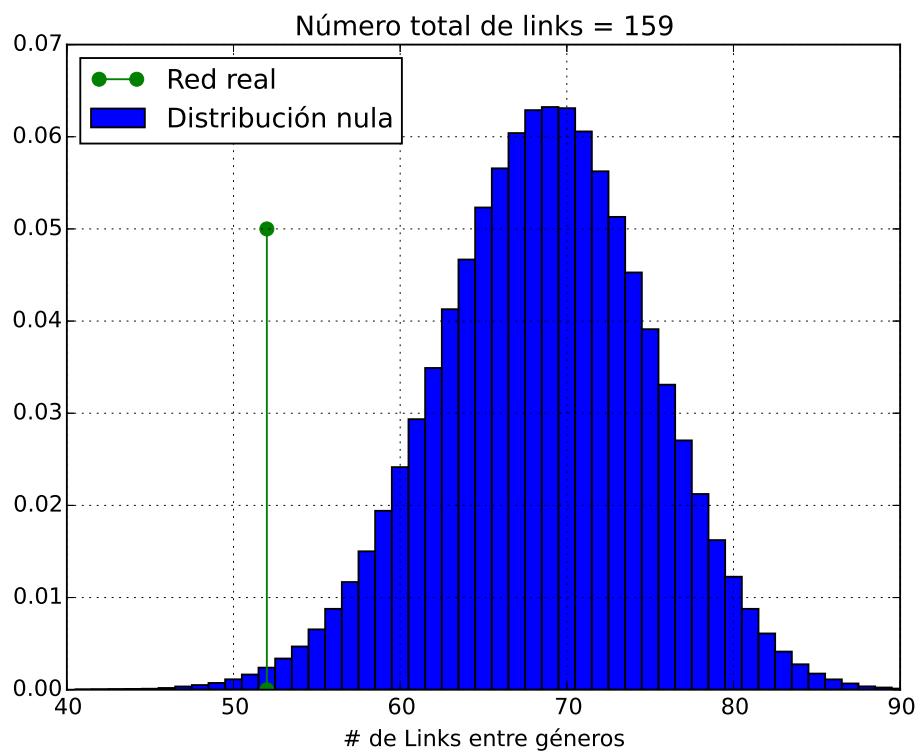


Figura 8: Histograma de links entre géneros sorteando sobre la topología de la red. En verde, cantidad de links entre géneros de la red actual (la altura de la barra no representa nada especial, es solo indicativa).

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

Como puede deducirse de la figura 8, 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:

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

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 9. Si lo comparamos con el caso de remover nodos al azar sin más criterio, es muy difícil obtener este resultado. En la figura 10, 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.

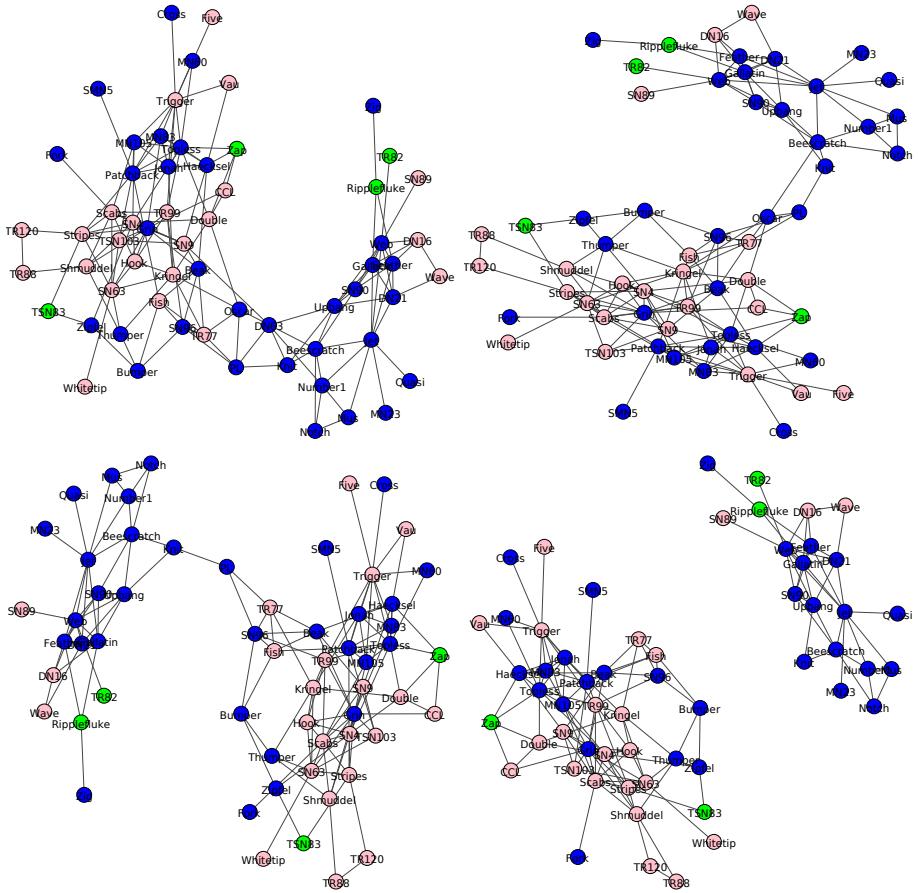


Figura 9: Layout de la red al remover el nodo con mayor betweenness: remoción de izquierda a derecha, y de arriba hacia abajo. Con solo remover 4 nodos, la red se descompone en dos subgrafos de tamaño comparable.

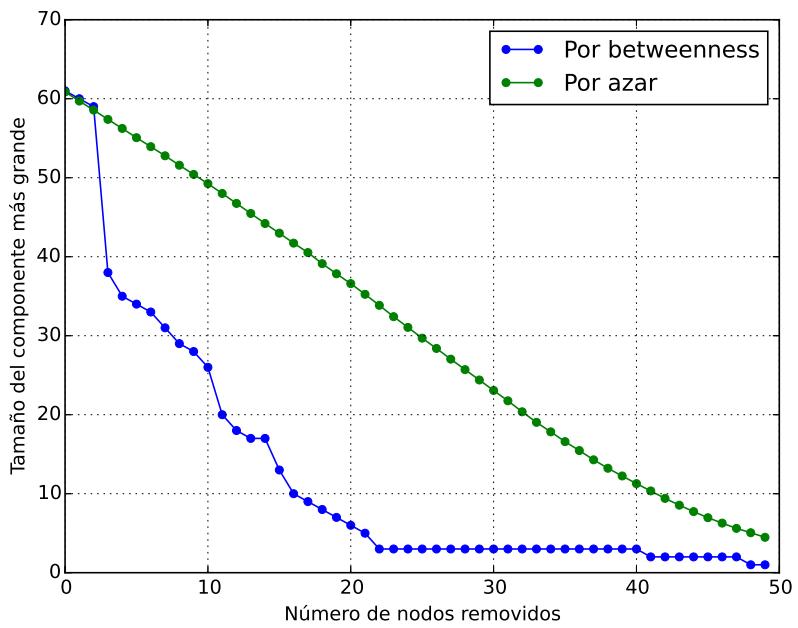


Figura 10: Tamaño del componente más grande a medida que se remueven nodos según betweenness, o al azar, respectivamente. Observar la caída abrupta al remover 4 nodos de mayor betweenness. La curva correspondiente a la forma azarosa es un promedio de 1000 configuraciones.

3. Ejercicio 3

En esta sección consideramos 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 librería `igraph` (versión 0.7.1) de Python. La distribución P_k de dichos valores de grado se observa en la figura 11, donde claramente observamos un fuerte comportamiento tipo power-law, a lo largo de tres 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 $p\text{-value}=0,5$, $k_{min}=3,9$, y del exponente $\alpha = 1,37$.

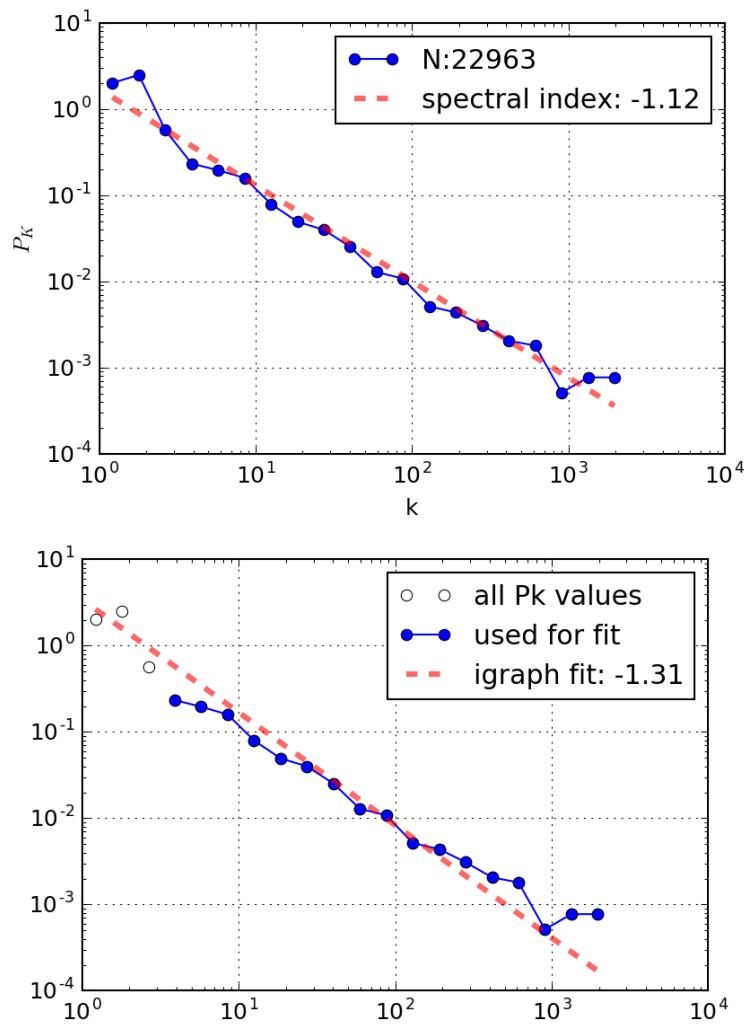


Figura 11: Distribución de probabilidad P_k de los valores de grado k de la red de Mark Newman. Observamos que la distribución tiene fuerte comportamiento tipo power-law, especialmente en el rango $k \sim (3 - 400)$.

4. Ejercicio 4

La assortatividad/homofilia es la tendencia de un nodo, de una red, se conecte con otros de características similares, esto en términos matemáticos es evaluar la correlación de links entre nodos del mismo tipo [Newman, 2003]. Esta propiedad tiende a ser característica para tipos de redes distintas: En redes de amistad por ejemplo se tiende a hacer uniones entre nodos *parecidos*, mientras que en redes biológicas, los *hubs* tienen a evadirse entre ellos y asociarse con nodos de menor grado [Newman, 2010, Barabási, 2016].

En este ejercicio se plantea evaluar la assortatividad de cuatro redes (dos sociales y dos biológicas) a través de dos métodos distintos: Mediante la estimación de la correlación de grado a partir del grado medio del vecindario de un nodo [Barabási, 2016]; y a través del Coeficiente de Correlación de Grado propuesto por Newman [Newman, 2003, 2010].

4.1. Consideraciones generales

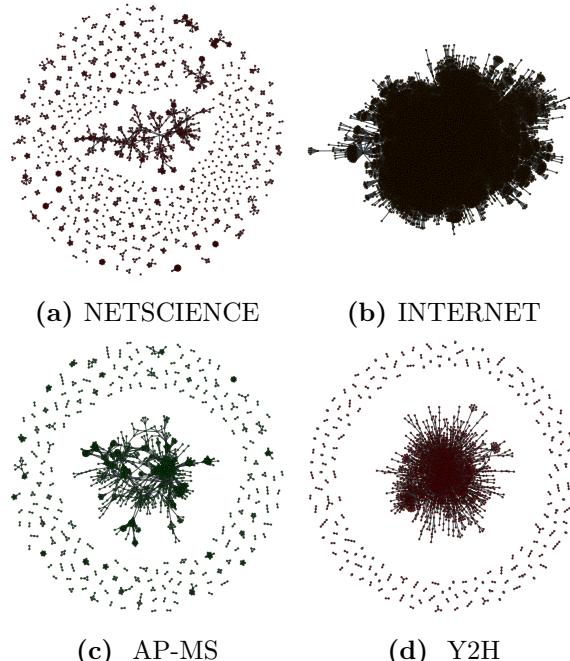


Figura 12: Layout SFDP para cada red estudiada
(script `plot.py <archivo> -f <fmt>`).

Consideremos las redes de colaboraciones científicas (NETSCIENCE),

red de internet (INTERNET) y dos redes de levadura analizadas en el ejercicio 1 (AP-MS y Y2H), las cuales son mostradas en la figura 12.

Un método alternativo para evaluar assortividad es analizar la matriz de correlación entre grados (*mixing matrix*) la cual se presenta en la figura 13.

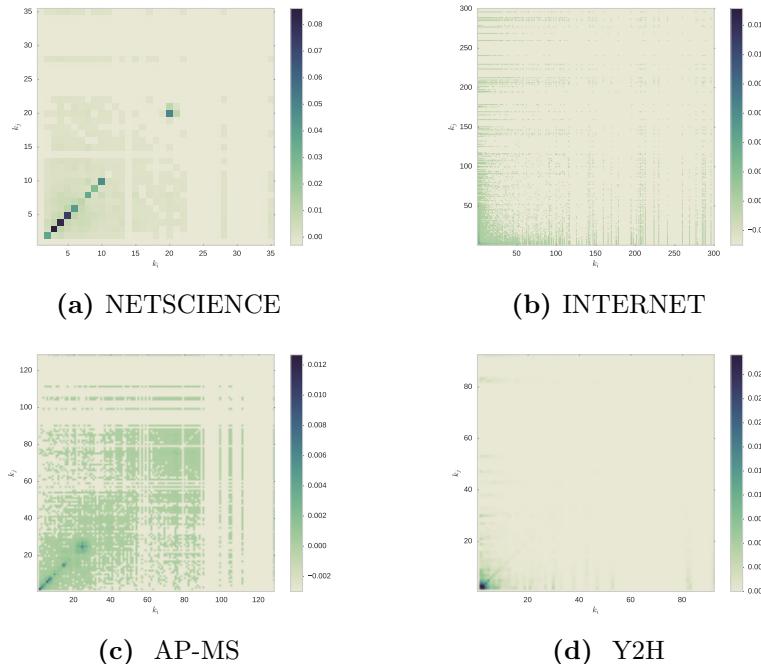


Figura 13: Mixing Matrix para cada caso. En la red INTERNET se ha ajustado los límites de gráfico de manera que sean visibles las correlaciones de bajo grado (script `degree_mixing_plot.py <archivo> -f <fmt>`)

De la figura se puede observar que, para todos los casos, la mayor densidad de correlación se puede encontrar en los grados pequeños, y para NETSCIENCE y AP-MS parece seguir un comportamiento lineal (*hubs* conectados con *hubs*), mientras que en las redes INTERNET y Y2H se puede observar que los nodos de alto grado tienden a conectarse con nodos de bajo grado. Sin embargo no es posible tener más que un valor estimativo de la assortividad con este método visual. Es necesario evaluar otras metodología.

4.2. Grado medio del vecindario k_{nn} (revising Barabási [2016])

En esta metodología estamos interesados en analizar el comportamiento de los vecinos de un nodo de grado k , para ello consideramos el grado medio

de los vecinos k_{nn} de un nodo de grado k , esto es

$$k_{nn}(k) = \sum_{k'} k' P(k'|k)$$

este promedio está hecho sobre la probabilidad de que: dado un nodo de grado k , tenga un vecino de grado k' . Para el caso de una red neutra, en que no existe correlacion entre k y k' ($\text{cov}(k, k') = 0$) se tiene que

$$P(k'|k) = P(k') \equiv q_{k'},$$

luego

$$k_{nn}(k) = \sum_{k'} k' q_{k'}$$

El modelo de Barabási-Albert propone que esta probabilidad es $q_{k'} = \frac{k' p'_k}{\langle k \rangle}$, la probabilidad de encontrar un nodo de grado k' al seleccionar un link aleatorio. Así, para el caso neutro

$$k_{nn}(k) = \frac{\langle k^2 \rangle}{\langle k \rangle},$$

es decir, es constante e independiente de k . Para el caso general, basandose en datos reales, se plantea el modelo

$$k_{nn}(k) \sim k^\mu. \quad (3)$$

A continuación se aplica el modelo power law a las redes mostradas en la figura 12 (ver figura 14). Aquí se pueden observar los comportamientos asortativos (NETSCIENCE y AP-MS) y disortativos (INTERNET y Y2H).

4.3. Coeficiente de Correlación de Grado

Por otro lado Newman [2010, 2003] propuso un coeficiente de correlación de caracteristica x (tambien llamado *coeficiente de asortividad*) dado por

$$r = \frac{\sum_{ij} (A_{ij} - k_i k_j / 2m) x_i x_j}{\sum_{ij} k_i \delta_{ij} - k_i k_j / 2m} \quad (4)$$

donde x es la caracteristica una característica del nodo, A la matriz de adyacencia, m el número total de links y k_i el grado del nodo i . En el caso

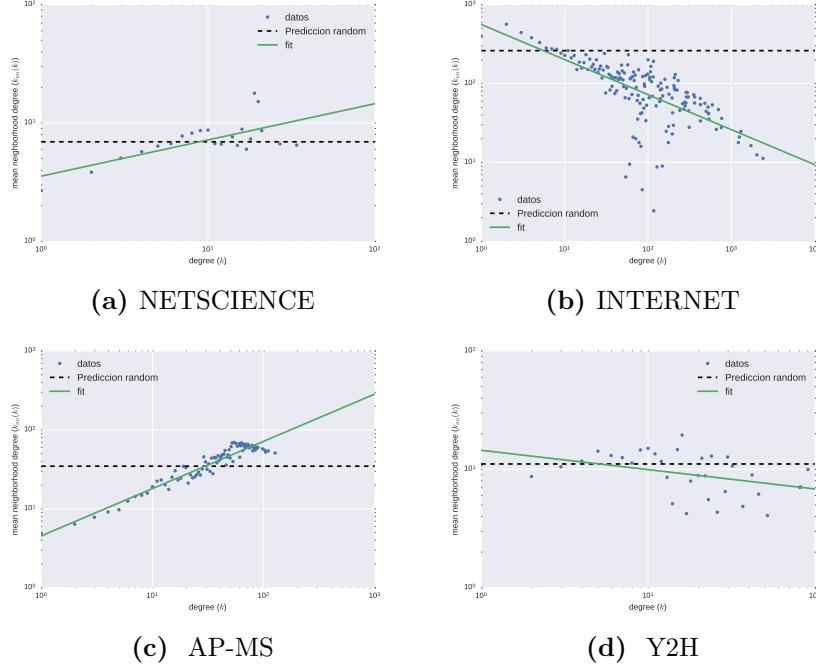


Figura 14: Grado medio del vecindario de k para cada red en escala log-log. Fiteo power law (Eq. 3) para cada caso: (a) $\mu \approx 0,306 \pm 0,071$, (b) $\mu \approx -0,444 \pm 0,040$, (c) $\mu \approx 0,599 \pm 0,019$, (d) $\mu \approx -0,163 \pm 0,064$ (script `assortivity_powerlaw.py <archivo> -f <fmt> -l`).

particular en que la característica de interés es el grado del nodo, entonces la Ec. 4 es

$$r = \frac{\sum_{ij} (A_{ij} - k_i k_j / 2m) k_i k_j}{\sum_{ij} (k_i \delta_{ij} - k_i k_j / 2m) k_i k_j} \quad (5)$$

Cabe notar que en Newman [2010] se recomienda implementar de manera distinta para evitar exceso de computos numéricos. Así, tomando $S_e = \sum_{ij} A_{ij} k_i k_j = 2 \sum_{\text{links}(i,j)} k_i k_j$, $S_1 = \sum_i k_i$, $S_2 = \sum_i k_i^2$ y $S_3 = \sum_i k_i^3$, y reemplazando en Ec. 5

$$r = \frac{S_1 S_e - S_2^2}{S_1 S_3 - S_2^2} \quad (6)$$

Una implementación del algoritmo descrito se puede encontrar en `assortative_newman.py`.

4.4. Discusión

Para las redes estudiadas, la Tabla II resume los resultados obtenidos

Tabla II: Tabla resumen de la asortatividad de las cuatro redes estudiadas.

Red	Tipo de Asortatividad	μ	r
NETSCIENCE	Asortativa	0.306	0.461
INTERNET	Disortativa	-0.444	-0.198
AP-MS	Asortativa	0.599	0.461
Y2H	Disortativa	-0.163	-0.041

Newman [2003] reporta las asortatividad de 27 redes y obtiene que ciertos rangos de asortatividad son caracteristicos del tipo de red en cuestión, por ejemplo, por un lado redes biológicas suelen ser disortativas y por otro las redes sociales tienen un comportamiento asortativo. En particular, en internet los servidores (*hubs*) no suelen conectarse entre ellos si no que más bien reciben un gigantesco número de usuarios (nodos) y servidores pequeños, lo que explica el comportamiento disortativo. Por otro lado, la red AP-MS está construida a partir de medición de afinidad, generando grandes clusters altamente conectados entre ellos, está manera de construir la red es probable que este asociada con la asortatividad reportada. Sin embargo, es necesario analizar rigurosamente tipo de asortatividad/disotatividad de estas redes (si son disortatividades estructurales o no). Los ultimos dos casos (NETSCIENCE y Y2H) responden a los comportamientos usuales para sus categorías (redes sociales y biológicas respectivamente).

Referencias

- Albert-László Barabási. *Network science*. Cambridge University Press, 2016.
- Sean R Collins, Patrick Kemmeren, Xue-Chu Zhao, Jack F Greenblatt, Forrest Spencer, Frank CP Holstege, Jonathan S Weissman, and Nevan J Krogan. Toward a comprehensive atlas of the physical interactome of *saccharomyces cerevisiae*. *Molecular & Cellular Proteomics*, 6(3):439–450, 2007.
- Mark Newman. *Networks: an introduction*. Oxford university press, 2010.
- Mark EJ Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.

Teresa Reguly, Ashton Breitkreutz, Lorrie Boucher, Bobby-Joe Breitkreutz, Gary C Hon, Chad L Myers, Ainslie Parsons, Helena Friesen, Rose Oughtred, Amy Tong, et al. Comprehensive curation and analysis of global interaction networks in *saccharomyces cerevisiae*. *Journal of biology*, 5 (4):1, 2006.

Haiyuan Yu, Pascal Braun, Muhammed A Yildirim, Irma Lemmens, Kavitha Venkatesan, Julie Sahalie, Tomoko Hirozane-Kishikawa, Fana Gebreab, Na Li, Nicolas Simonis, et al. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, 2008.