Teoría de Algoritmos II

Parcialito 1

Autor: Alejo Tomás Mariño
Padrón: 105645

<u>Consigna</u>:

1. Determinar:
a. El diámetro de la red.
b. El grado promedio de la red.
c. El coeficiente de clustering promedio de la red.

2. Indicar si existe algún tipo de homofilia y qué tipo de homofilia es. Si no hay homofilia por ningún criterio, explicar. Justificar detalladamente.

3. Determinar los puentes (globales o locales) en dicha red.

4.
a. Determinar un tipo de centralidad que podría ser útil calcular para esta red, justificando.
b. Realizar una representación gráfica de dicha red, considerando la centralidad de los distintos países dada por la métrica del punto a (tamaño de los nodos proporcional a dicha métrica).

5.
a. Obtener una simulación de un modelado de Erdös-Rényi que corresponda a los parámetros de esta red.
b. Obtener una simulación de un modelado de Preferential Attachment (ley de potencias) que corresponda a los parámetros de esta red.
c. Obtener una representación de anonymous walks tanto de la red original como para las dos simuladas en los puntos a y b. Determinar por distancia coseno cuál sería la simulación más afín.

1. a. El diámetro de la red dada, es decir, la distancia más corta entre los dos nodos más distantes de la red, según gephi (el workspace de gephi con el grafo generado por el .csv dado estará adjuntado en el .zip entregado junto con el codigo) es 5 con una distancia promedio de 2.247.

Network Diameter 5 Run

b. El grado promedio de la red no es más que el promedio de los grados de cada uno de los vértices en la red, siendo el grado de un vértice para un grafo no dirigido la cantidad de aristas del vértice. Según gephi, el grado de la red es 24,9

Average Degree 24.908 Run

c. El coeficiente de clustering de un vértice en sí no es más que, teniendo un vértice A la probabilidad de que dos vértices adyacentes B y C de A sean adyacentes entre sí. El coeficiente de clustering promedio no es más que la sumatoria de todos los coeficientes de clustering individuales / el número de vértices del grafo. Según gephi, el coeficiente de clustering de la red es 0.687.

```
Average Clustering Coefficient: 0.687
Total triangles: 21087
```

2. Para ver si hay homofilia en la red voy a dividir a los países de la red en dos grupos, Países “Ricos” (Aquellos con World PPP GDP per capita >= 200%) y países “Pobres” (Aquellos con World PPP GDP per capita < 200%).
Source: https://www.worldometers.info/gdp/gdp-per-capita/
Data de la red:

```
----------Rich vs Poor countries proportions----------

Number of nodes in given network: 229
Number of edges in given network: 2852
Proportion of Rich countries in the network: 0.1615720524017467
Proportion of Poor countries in the network: 0.8384279475982532
Proportion of edges that have a Rich country on one end, and a Poor country on its other: 0.4642356241234222
Proportion of Rich countries' edges which connect a Poor and a Rich country: 0.623352165725047
Proportion of Poor countries' edges which connect a Poor and a Rich country: 0.36983240223463687
```

Ahora, para evaluar si hay o no homofilia en la red…
Llamemos A a la probabilidad de que un nodo cualquiera de la red sea un país “Rico”, sabemos que A ~= 0,1615 (~16,15%) y B a la probabilidad de que un nodo cualquiera de la red sea un país “Pobre”, sabemos que B ~= 0,8384 (~84,85%).
La probabilidad de que dos vértices cualesquiera que estén unidos por una arista sean ambos ricos es A^2 ~= 0,0261 (~2,61%) y para que ambos sean pobres es B^2 ~= 0,7029 (~70,29%). La probabilidad de que una arista tenga en uno de sus extremos un país rico y en el otro extremo un país pobre es 2*A*B ~= 0,2709 (~27,09%).
Al ver la cantidad de aristas en la red que “cruzan” de un país rico a uno pobre, podemos observar que el valor (~0,4642 (~46,42%)) es claramente más alto que el teórico.
Haciendo la cuenta: (proporción real de aristas que cruzan) / (proporción teórica de aristas que cruzan) ~= 1.7134 (~171,34%).
Con esto se puede observar de que claramente NO hay homofilia en esta red, tomando como criterio el elegido para la distinción entre países “Pobres” y “Ricos”.
Algunas conclusiones que se pueden sacar de esto son:

Algo que cabe destacar es que por mas que no haya homofilia, solo un ~37% de las aristas que tienen al menos 1 país pobre en sus extremos conectan con un país rico en

el otro extremo, lo que uno lo puede acreditar al menor poder adquisitivo de los ciudadanos de los países pobres, haciendo que el turismo mediante viajes hacia países ricos no sea una opción tan viable, algo a notar de lo cual me di cuenta analizando la data proporcionada y pensando el porqué de los números es que si una persona de un país rico ve valor en viajar a un país pobre (ya sea por cuestiones laborales, religiosas u otras), lo hará, creando esta arista del país pobre al rico. Lo que puede confundir y enturbiar la realidad en estos datos llevándonos a conclusiones erróneas es que puede que de estos países pobres, jamás se hagan viajes HACIA el país rico y simplemente sirvan para que aquellos ciudadanos de países ricos puedan visitar el país pobre. Esto nos hace repensar ese 37% de aristas que comparten con países ricos, siendo una absoluta posibilidad el hecho de que los viajes que se hagan desde países pobres hacia los más ricos sea mucho menor de lo que la cifra puede hacer pensar.
Contrastando, las aristas que contienen al menos 1 país rico en sus extremos, las cuales tienen una proporción de ~62% de cruce con un país no pobre, mostrando claramente que del total de los viajes de los países ricos, casi 2 tercios de los mismos son hacia países pobres. Algunas razones sobre las cuales especulo que tienen relevancia pueden ser el turismo a países más pobres, los cuales ofrecerán más valor por su dinero que si decidiese hacer turismo en un país que también es rico. Otro motivo podría ser por visitas familiares, de estudiantes/trabajadores que hayan conseguido algún tipo de responsabilidad/cargo laboral en un país rico y vuelvan a su país de origen pobre para visitar a su familia. También podría entrar el juego viajes de compras que hagan ciertos individuos desde países ricos hacia pobres, para gastar su dinero en un país el cual, ya sea por inflación de la moneda local u otros factores económicos, permite al viajante obtener más valor por su moneda natal.

En conclusión, no hay homofilia en la red dada por el mero hecho de que dividir al mundo en “Ricos” y “Pobres” mediante el índice proporcionado (al menos con el cut off point de 200% usado) no tiene gran sentido/utilidad a la hora de analizar los viajes entre países por razones dadas previamente, o más que nada por el hecho de que el mundo en el cual vivimos hoy en día está altamente globalizado, lo que en consecuencia lleva a que básicamente la gran mayoría de los países, aun aquellos considerados como pobres en la métrica dada tiene algún tipo de aeropuerto que actúe como receptor de turismo global, y si existe la posibilidad de turismo, siempre habrá un viajante en un país rico con ingreso disponible suficiente como para viajar a un país pobre, lo que terminará generando una arista.
En comparación, una métrica binaria para dividir la red a la cual le veo un cierto nivel de utilidad y en la cual es probable que haya homofilia sería dividir a la red en países como Cuba, Corea del Norte, Rusia, etc (países con influencia comunista/pseudo-comunista) y el resto del mundo, para ver si realmente se dan cruces suficientes para superar o no el margen de homofilia teórico.

Haciendo un rapido analisis de esta situación, usando solo a Corea del Norte y Cuba (2 países prácticamente “aislados” del resto del mundo) obtengo los siguientes datos:

```
----------Red countries vs rest of the world proportions----------

Number of nodes in given network: 229
Number of edges in given network: 2852
Proportion of red countries in the network: 0.008733624454148471
Proportion of not red countries in the network: 0.9912663755458515
Proportion of edges that have a red country on one end, and a not red country on its other: 0.012622720897615708
Proportion of red countries' edges which connect a not red and a red country: 1.0
Proportion of not red countries' edges which connect a not red and a red country: 0.006351446718419196
```

Proporción de cruces teóricos: 0,01731469651608474113739249823611
Proporción de cruces reales: 0,012622720897615708
Por lo que: reales/teóricos = 0,7290 (~72,90%)
Concluyo que para esta división binaria de la red hay homofilia, la razón está dada no más que por el criterio en sí (países “aislados”).

3. La cantidad de puentes globales de la red (aristas que al ser removidas aumentan la cantidad de componentes conexas en la red) es 9. Los mismos son:

```
Puente 1: 'Fiji'-'Tuvalu'
Puente 2: 'United States'-'American Samoa'
Puente 3: 'United Kingdom'-'Saint Helena'
Puente 4: 'Canada'-'Saint Pierre and Miquelon'
Puente 5: 'Antigua and Barbuda'-'Montserrat'
Puente 6: 'New Zealand'-'Niue'
Puente 7: 'South Africa'-'Lesotho'
Puente 8: 'South Africa'-'Swaziland'
Puente 9: 'Burma'-'Myanmar'
```

La cantidad de puentes locales en la red (aristas las cuales unen 2 partes de la red las cuales no tendrían otra conexión si no existiesen) es 11. Los mismos son:

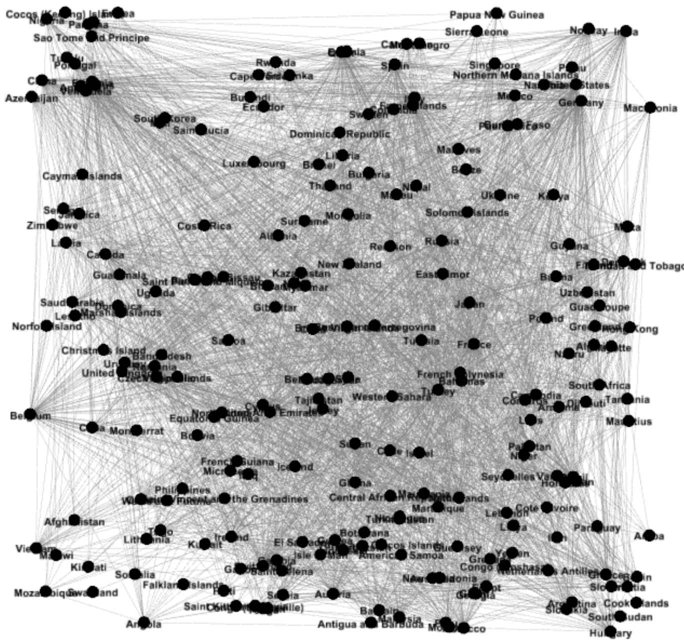
```
----------Puentes Locales----------

Puente 1: 'Papua New Guinea'-'Micronesia'
Puente 2: 'Fiji'-'Tuvalu'
Puente 3: 'Micronesia'-'Marshall Islands'
Puente 4: 'United States'-'American Samoa'
Puente 5: 'United Kingdom'-'Saint Helena'
Puente 6: 'Canada'-'Saint Pierre and Miquelon'
Puente 7: 'Antigua and Barbuda'-'Montserrat'
Puente 8: 'New Zealand'-'Niue'
Puente 9: 'South Africa'-'Lesotho'
Puente 10: 'South Africa'-'Swaziland'
Puente 11: 'Burma'-'Myanmar'
```

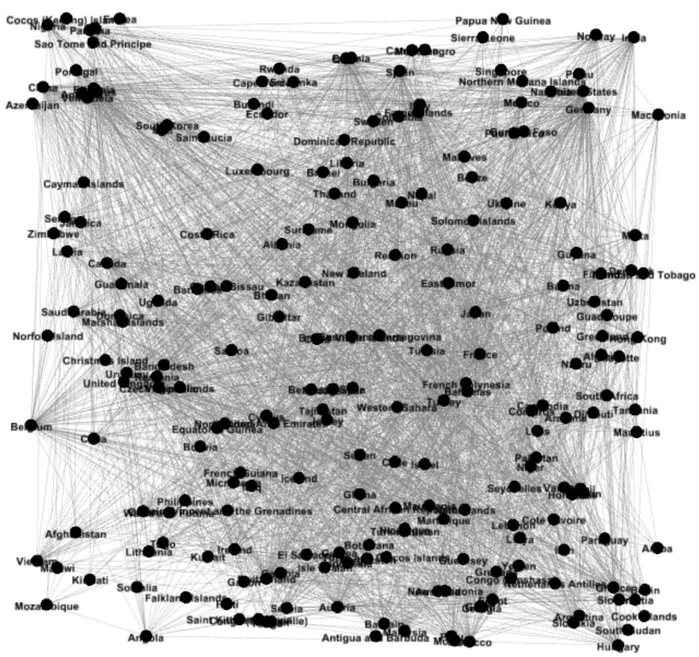
Biblioteca usada: https://networkx.org/documentation/stable/reference/algorithms/bridges.html

4. a. Inicialmente pienso que k-core decomposition vendría bien, de esa manera los nodos con mayor cantidad de viajes (uno lo podría pensar como los que tienen más influencia global/cultural) quedarán más “grandes” en una representación gráfica y mientras más nos alejemos del medio de la red, la influencia global de los países irá disminuyendo. Harmonic closeness centrality también probablemente funcione bien.

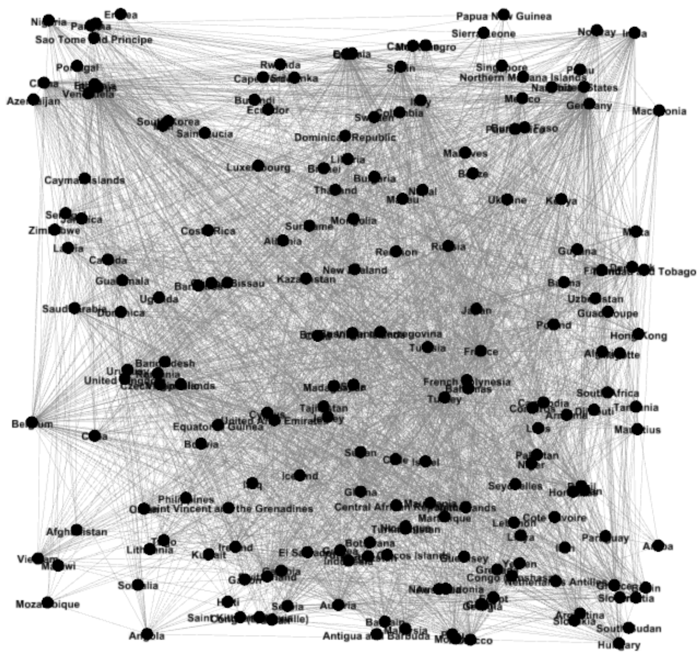
b. No encontré forma de que el gráfico de networkx de k-core decomposition se visualice de una manera que me ayude a sacar conclusiones por lo que lo hice con gephi, no tengo un gran manejo de las tools por lo que no se realmente como haria para graficar los del core más grande en el medio y más grandes y los más pequeños en cores más “externos”.
Nota para Martin: si podes recomendarme documentación o alguna forma de aprender a graphear (como las redes en los slides de las clases) los distintos tipos de diccionarios (conjuntos de nodos) que devuelven las funciones de NetworkX como closeness_centrality() o harmonic_centrality() cambiando el tamaño de los nodos según su valor te lo agradeceria mil veces, más que nada porque odio no poder sacar conclusiones porque no se manejar el software apropiadamente.
A continuación, muestro los k-cores para distintos k:
k=1



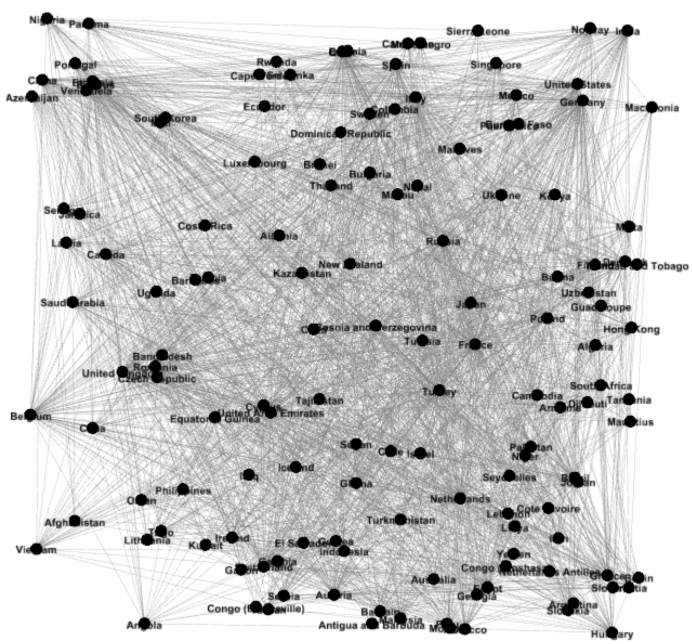
k=2



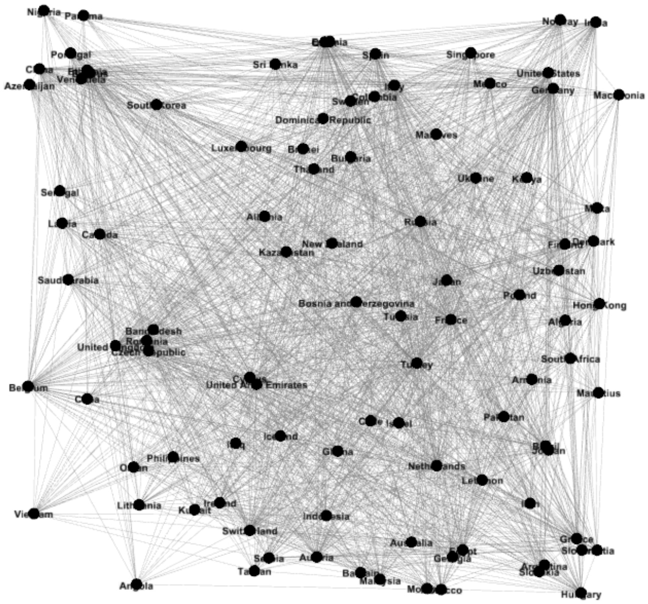
k=5


Papua New Guinea
Sierra Leone
Norway
India
Sao Tome and Principe
Portugal
Rwanda
Spain
Singapore
Northern Mariana Islands
Azerbaijan
Burundi
Ecuador
Mexico
Germany
Macedonia
South Korea
Saint Lucia
Dominican Republic
Burkina Faso
Luxembourg
Bulgaria
Cayman Islands
Thailand
Ukraine
Kenya
Zimbabwe
Costa Rica
Mongolia
Suriname
Russia
Guyana
Trinidad and Tobago
New Zealand
Kazakhstan
Guatemala
Uganda
Saudi Arabia
Japan
Uzbekistan
Guadeloupe
Poland
Hong Kong
Tunisia
France
French Polynesia
Bangladesh
United Kingdom
Czech Republic
South Africa
Tanzania
Belgium
Tajikistan
Equatorial Guinea
United Arab Emirates
Mauritius
Sudan
Iceland
Pakistan
Niger
Seychelles
Philippines
Saint Vincent and the Grenadines
Central African Republic
Martinique
Afghanistan
Cote d'Ivoire
Paraguay
Lithuania
Kuwait
Botswana
Aruba
Mozambique
Somalia
Angola
Antigua and Barbuda
Netherlands Antilles
South Sudan
Hungary

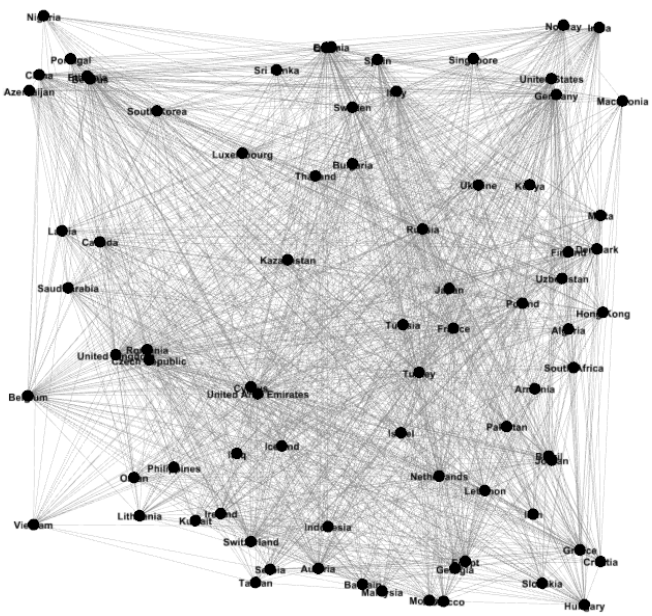
k=10


Sierra Leone
Norway
India
Portugal
Rwanda
Spain
Singapore
United States
Germany
Macedonia
South Korea
Ecuador
Dominican Republic
Burkina Faso
Luxembourg
Bulgaria
Thailand
Ukraine
Kenya
Costa Rica
Russia
New Zealand
Kazakhstan
Trinidad and Tobago
Uganda
Saudi Arabia
Japan
Uzbekistan
Guadeloupe
Poland
Hong Kong
Bosnia and Herzegovina
Tunisia
France
Bangladesh
Czech Republic
South Africa
Tanzania
Belgium
Turkey
Tajikistan
Equatorial Guinea
United Arab Emirates
Mauritius
Sudan
Iceland
Pakistan
Niger
Seychelles
Philippines
Ghana
Netherlands
Cote d'Ivoire
Afghanistan
Turkmenistan
Lithuania
Kuwait
Indonesia
Vietnam
Australia
Georgia
Netherlands Antilles
Angola
Antigua and Barbuda
Hungary

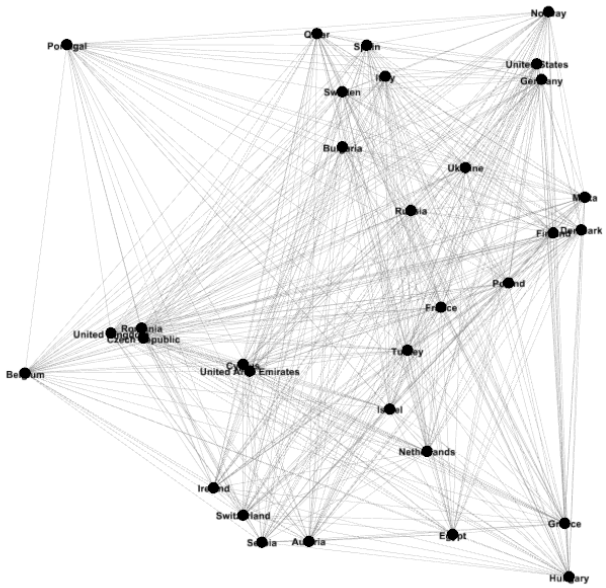
k=15


South Korea
Sri Lanka
Dominican Republic
Luxembourg
Singapore
United States
Germany
Macedonia
Portugal
Azerbaijan
New Zealand
Kazakhstan
Bosnia and Herzegovina
Saudi Arabia
Canada
Senegal
Uzbekistan
Hong Kong
South Africa
United Arab Emirates
Philippines
Lithuania
Switzerland
Netherlands
Indonesia
Australia
Hungary
Turkey
France
Russia
Bulgaria
Ukraine
Mauritius

k=20


South Korea
Sri Lanka
Luxembourg
Singapore
United States
Germany
Macedonia
Portugal
Azerbaijan
Kazakhstan
Saudi Arabia
Uzbekistan
Hong Kong
South Africa
United Arab Emirates
Philippines
Lithuania
Switzerland
Netherlands
Indonesia
Hungary
Turkey
France
Russia
Bulgaria
Ukraine
Slovakia

k=27



Obviamente que si manejase alguna buena herramienta para graficar la k-core decomposition tendría a los nodos que se ven en k=27 en el core más alto, en el medio del gráfico, y a medida que voy disminuyendo el k-core y aristas nuevas van apareciendo, esas aristas irían en su respectivo k-core, alejándose de a poco del centro del gráfico.

5. a. Datos de la simulación del modelo Erdos-Renyi con los parámetros de la red original:

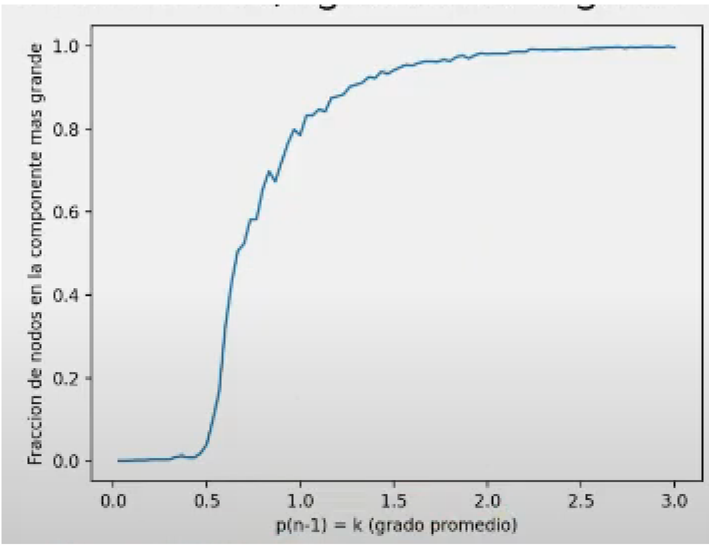
```
----------Erdos-Renyi simulation----------

Average degree of original network: 24.90829694323144
Average degree of Erdos-Renyi simulation: 25.100436681222707
Diameter of original network: 5
Diameter of Erdos-Renyi simulation: 3
Clustering coefficient of original network: 0.6601565365859736
Clustering coefficient of Erdos-Renyi simulation: 0.1133974483892974
Average degree divided by number of nodes (which should be close to the Erdos-Renyi clustering coefficient): 0.10876985564729887
```

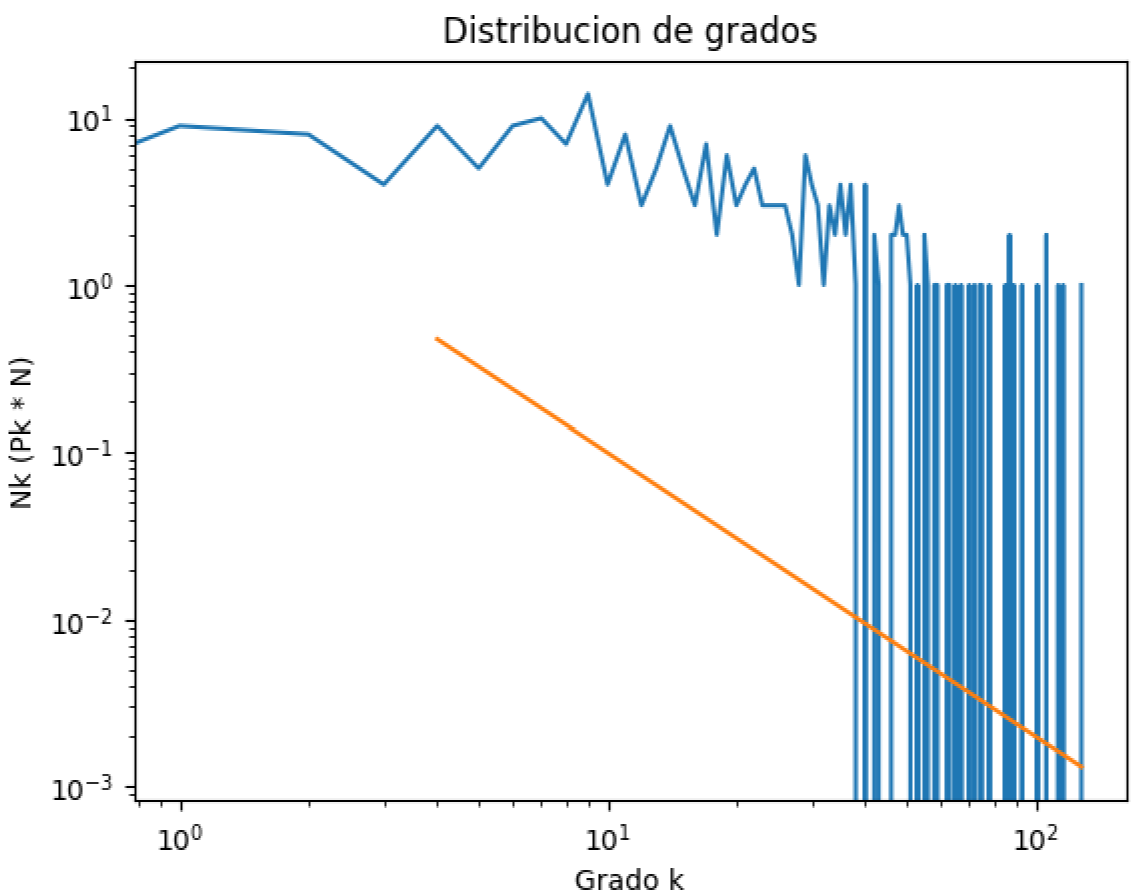
Podemos ver que el diámetro del E-R parece seguir la teoría ("Un grafo E-R puede crecer mucho en n pero sus nodos se mantendrán a una distancia relativamente cercana").

Vemos que el coeficiente de clustering del E-R, se asemeja tal como dice la teoría a k/n (con k el grado promedio del grafo original y n la cantidad total de nodos del grafo original) y es muy bajo comparado con el de la original como se esperaba.

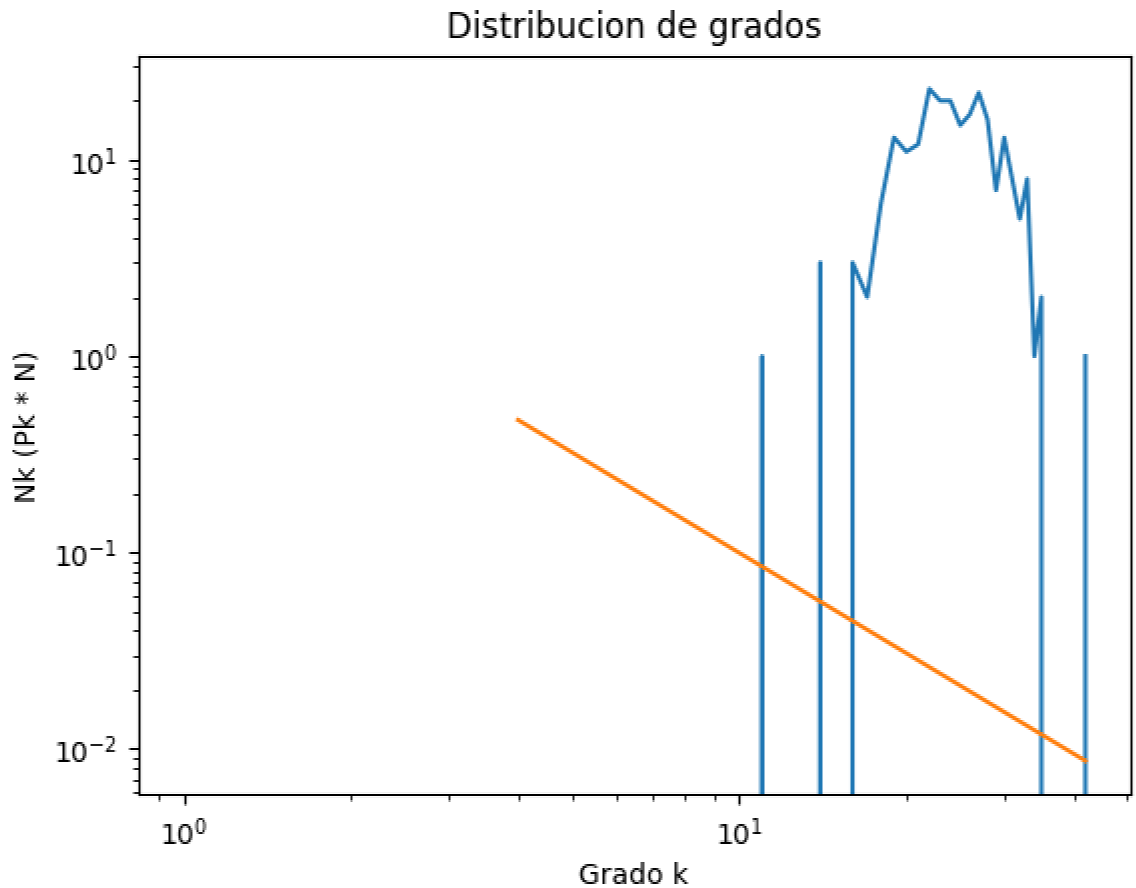
No es sorpresa que el E-R tenga una sola componente conexa ya que según este gráfico visto en clase es claro que con un grado promedio >24 no hay dudas que va a haber una única componente conexa:



Usando las funciones de graficación proporcionadas (perdón pero no conocía otra forma de graficar para graficar en escala normal), el resultado de ambas redes es:
Original:



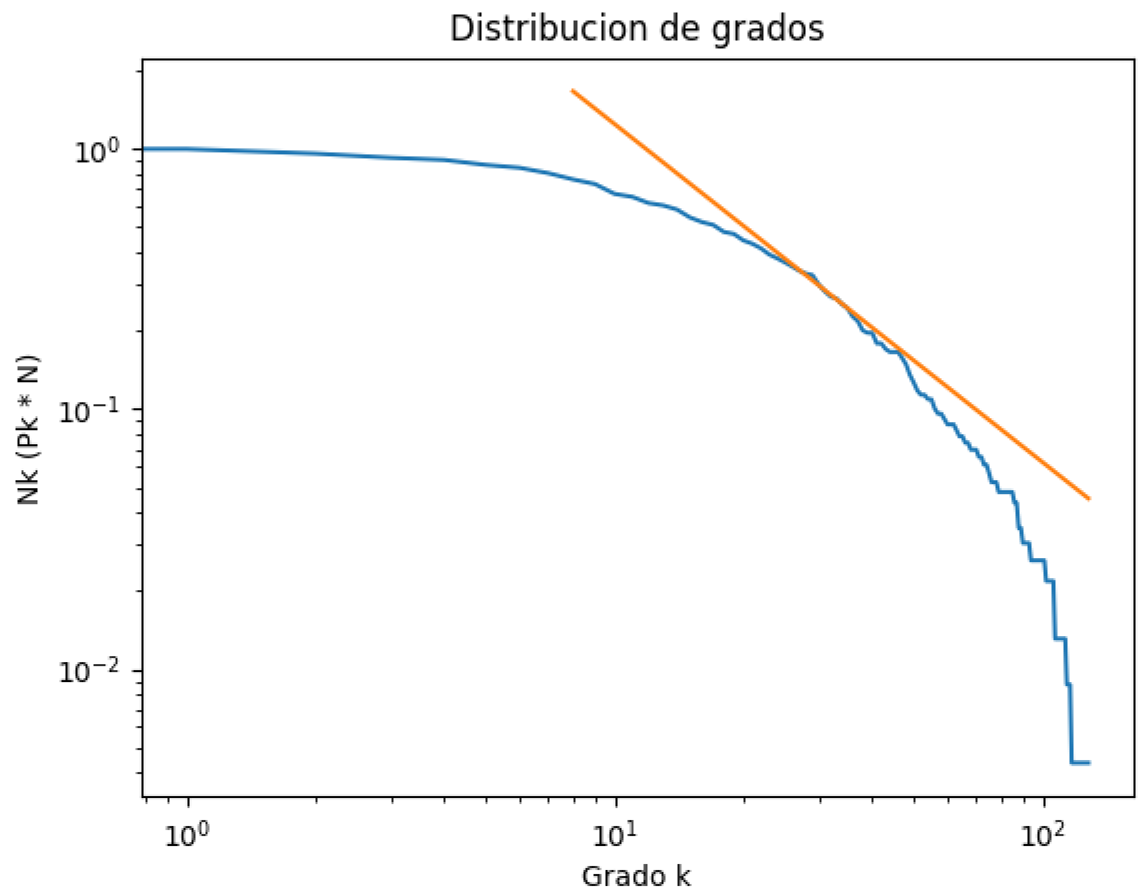
E-R:



Se puede ver que la distribución de grados de E-R parecería ser una campana de Gauss que es lo esperado.

Concluyendo, el modelo de Erdos-Renyi, tal como visto en clase no es muy bueno a la hora de replicar la distribución de grados de la red original, o el coeficiente de clustering de la misma. Aun así, imita la cantidad de componentes conexas satisfactoriamente y es capaz de mantener un diámetro bajo (similar a la red original), y aproxima medianamente bien las distancias de la red original.

b. Para el modelo de Preferential Attachment comencé calculando el alfa de mi red original usando ccdf y la función proporcionada, obteniendo el siguiente gráfico:



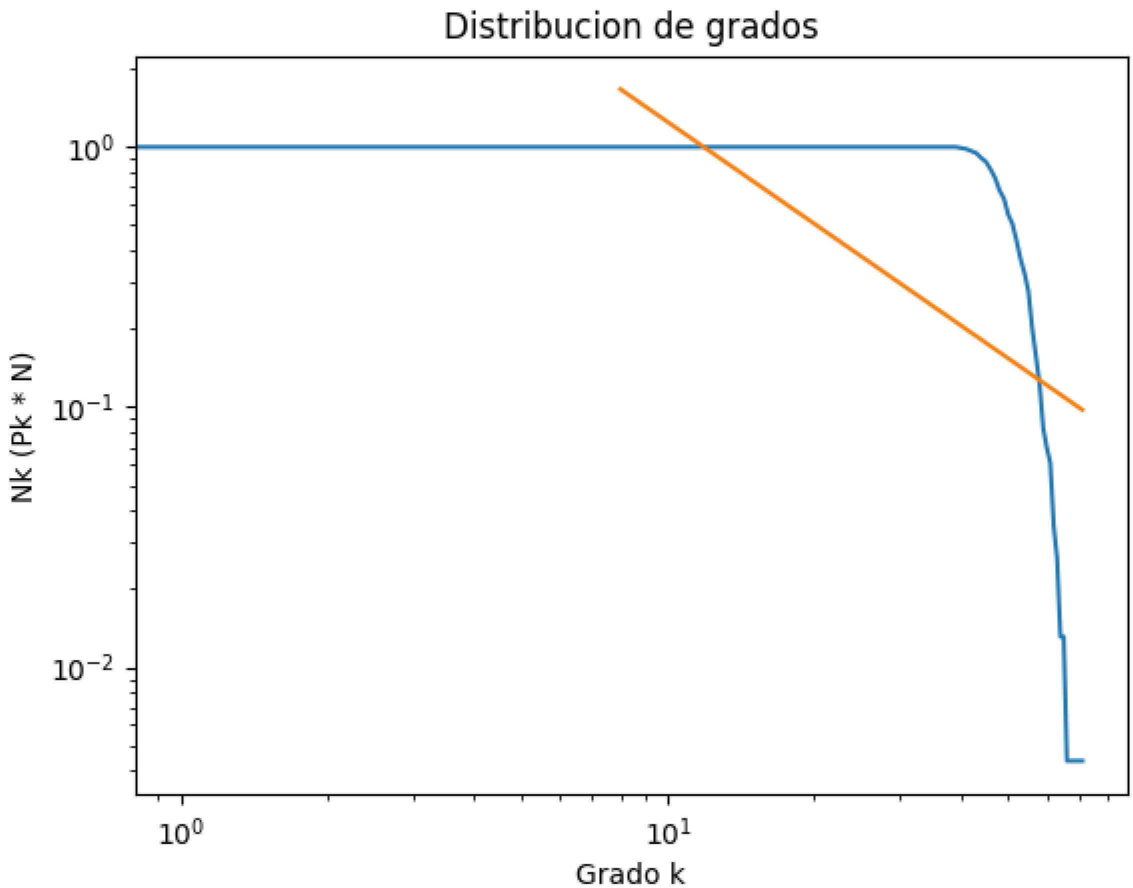
El gráfico fue resultado de un alfa = 2,3 pero usando máxima verosimilitud me dio alrededor de 2,18 por lo que voy a tomar 2,25 como mi valor de alfa para mi red original.

Sabemos que el valor de p se calcula de la siguiente manera:

$$p = 1 - \frac{1}{\alpha - 1}$$

Para mi alfa = 2,25, p = 0.2.

Graficando la distribución de grados de la red por preferential attachment se ve claramente como la gran mayoría de los nodos tienen un grado extremadamente alto, como se esperaba:



Otros datos:

```
Average degree of original network: 24.90829694323144
Average degree of Preferential Attachment simulation: 51.056768558951966
Diameter of original network: 5
Diameter of Preferential Attachment simulation: 2
Clustering coefficient of original network: 0.6601565365859736
Clustering coefficient of Preferential Attachment simulation: 0.22264306019431013
Average path length in original network: 2.247069639163411
Average path length in Preferential Attachment simulation: 1.7760668045660002
Connected components in original network: 1
Connected components in Preferential Attachment simulation: 1
```

Parece cumplirse que la distancia promedio para alfa entre 2 y 3 es log(log(n)).
Y es obvio que la cantidad de componentes de preferential attachment iba a ser 1.

c. Para los anonymous walks tomo L=7, uso el epsilon (0,1) y delta (0,01) dados por la función suministrada. Llamo a la función anonymous_walks() con cada uno de los 3 grafos y luego comparo los vectores 2 veces, una con el anonymous walk del original y

el anonymous walk del E-R y otra con el anonymous walk del original y el anonymous walk del preferential attachment.
Al ejecutar esto obtuve (lo corrí un par de veces para asegurarme de que no haya sido un caso borde):

```
----------Anonymous walk simulation----------

Cosine distance between original network and  Erdos-Renyi simulation: 0.00047098250163535305
Cosine distance between original network and Preferential Attachment simulation: 0.0020179370062819446
```

Lo que me lleva a concluir que la simulación de la red mediante Erdos-Renyi es superior (se asemeja más a la red original) a la de Preferential Attachment, algo que no esperaba (y sinceramente puede ser que haya resuelto de manera inadecuada, si fuese el caso me gustaria saber que hice mal).