



**FACULTAD
DE INGENIERIA**

Universidad de Buenos Aires

TEORÍA DE ALGORITMOS II

1° CUATRIMESTRE 2022

Red de Vuelos

AUTOR

Giampieri, Leonardo
<lgampieri@fi.uba.ar>

- #102 358

DOCENTE

Buchwald, Martín

Índice

1. Introducción	2
2. Diámetro, Grado Promedio, Coeficiente de Clustering	3
3. Homofilia	4
4. Puentes Locales y Globales	5
5. Centralidad	6
6. Modelados	7
6.1. Erdős-Rényi	8
6.2. Preferential Attachment	10
6.3. Anonymous Walks	11

1. Introducción

En el siguiente informe se presenta la resolución del 1er parcialito domiciliario, en el cual se trabaja sobre la siguiente [red de vuelos](#). La resolución en código se puede encontrar en el siguiente [repositorio](#). Como herramientas, se utilizó tanto Networkx como también Gephi.

2. Diámetro, Grado Promedio, Coeficiente de Clustering

Diámetro	Grado Promedio	C. de Clustering
5	24.91	0.66

Vemos que 66 % de las aristas forman parte de un triángulo, lo cual muestra que hay una tendencia alta a la clausura transitiva. Además, siguiendo con propiedades comunes de redes sociales, el diámetro es chico.

3. Homofilia

Para calcular si existe algún tipo de homofilia, elegí caracterizar a los países por el continente al que pertenecen, haciendo uso de las bibliotecas [pycountry](#) y [pycountry-convert](#), además de las funcionalidades ya programadas en la biblioteca de [homofilia](#).

Se obtuvo que el threshold para que no exista homofilia por continente en la red es de 0.81. Esto se obtuvo de la siguiente tabla de proporciones (separó América, ya que esto es lo que hace internamente pycountry):

Continente	Proporción de países
Europa	0.20
África	0.25
América del Norte	0.16
América del Sur	0.06
Asia	0.23
Oceanía	0.10

Esto significa que para que no exista homofilia, el 81 % de los vuelos deberían cruzar de continente, o lo que es lo mismo, por cada país, el 81 % de sus vuelos debería ser a países en otro continente. En la red sin embargo, la proporción de vuelos que cruzan continente es de 0.44. Si bien realizando los cálculos por continente, uno puede encontrar que los vuelos intercontinentales superan la cantidad de vuelos intracontinentales, en un estudio macroscópico, podemos ver que eso no es así. Hay 1242 vuelos que cruzan continente, de un total de 2852 vuelos.

Continente	Vuelos Intercontinentales	Vuelos Intracontinentales
Europa	857	563
África	406	325
América del Norte	273	168
América del Sur	134	45
Asia	737	461
Oceanía	48	77

Cabe destacar que la cantidad de vuelos intracontinentales a nivel de red se debería dividir por dos, ya que en esta tabla hay vuelos intercontinentales repetidos por cada par de continentes (debido a que la red es no dirigida).

Esto es lógico! No es posible desde cualquier país del mundo llegar a cualquier otro país de otro continente, que es lo que no tendría en cuenta un modelo aleatorio sobre la red. Este modelo es aquel que presupone el cálculo del threshold, ya que se asignarían aristas aleatoriamente para que se cumpla a nivel global con las proporciones de la red original. La homofilia en esta red nos está mostrando en cierta forma una realidad geográfica embebida en la misma.

4. Puentes Locales y Globales

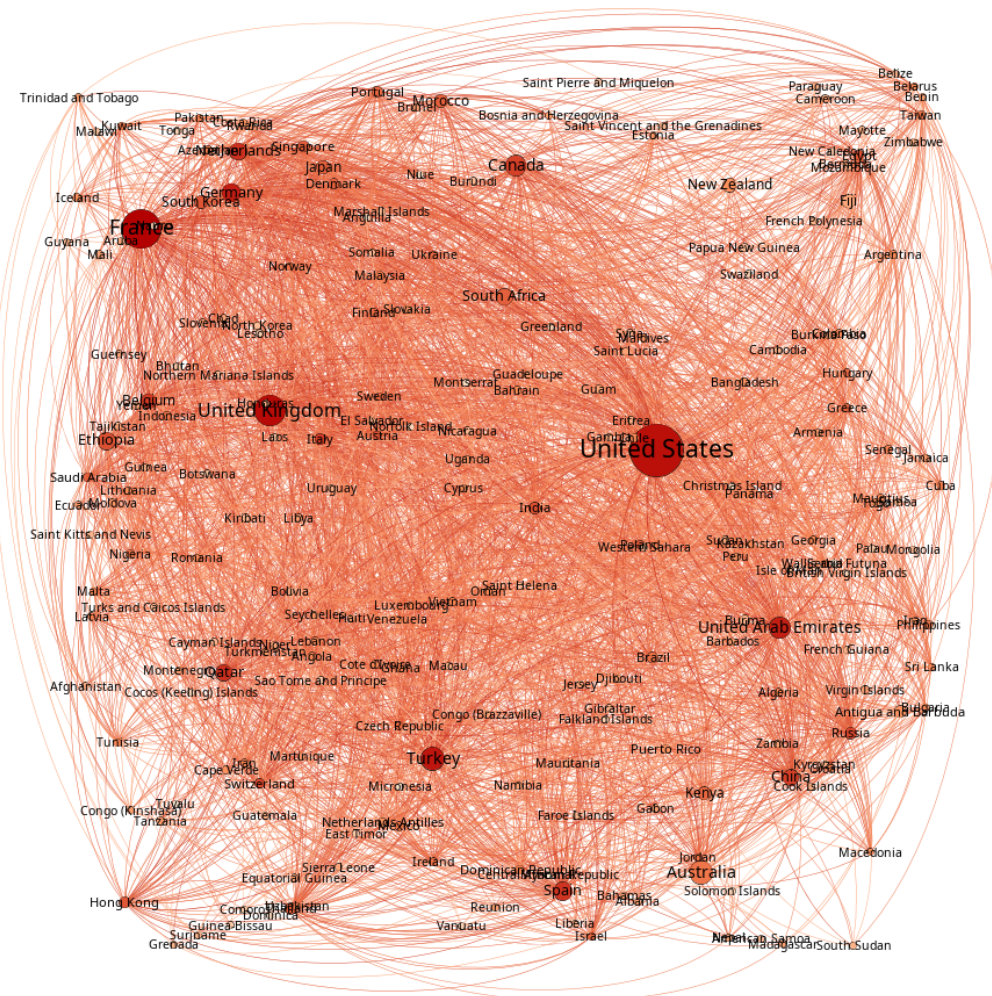
Globales
Fiji - Tuvalu
United States - American Samoa
United Kingdom - Saint Helena
Canada - Saint Pierre and Miquelon
Antigua and Barbuda - Montserrat
New Zealand - Niue
South Africa - Lesotho
South Africa - Swaziland
Burma - Myanmar

Como todos los puentes globales también son puentes locales, se repiten algunas aristas en la tabla.

Locales
Fiji - Tuvalu
United States - American Samoa
United Kingdom - Saint Helena
Canada - Saint Pierre and Miquelon
Antigua and Barbuda - Montserrat
New Zealand - Niue
South Africa - Lesotho
South Africa - Swaziland
Burma - Myanmar
Papua New Guinea - Micronesia
Micronesia - Marshall Islands

5. Centralidad

Elegí calcular Betweenness Centrality (de ahora en mas BC). Se podría pensar que un camino mínimo en la red de vuelos, para un par de países, nos devuelve el camino con menor cantidad de escalas. Si bien esto es cierto para la red, en la realidad no significa que sea el vuelo mas corto, simplemente es un posible vuelo con ciertas escalas. De todas maneras, los países con mayor puntaje de BC serán aquellos que en mayor cantidad de escalas sean intermediarios.

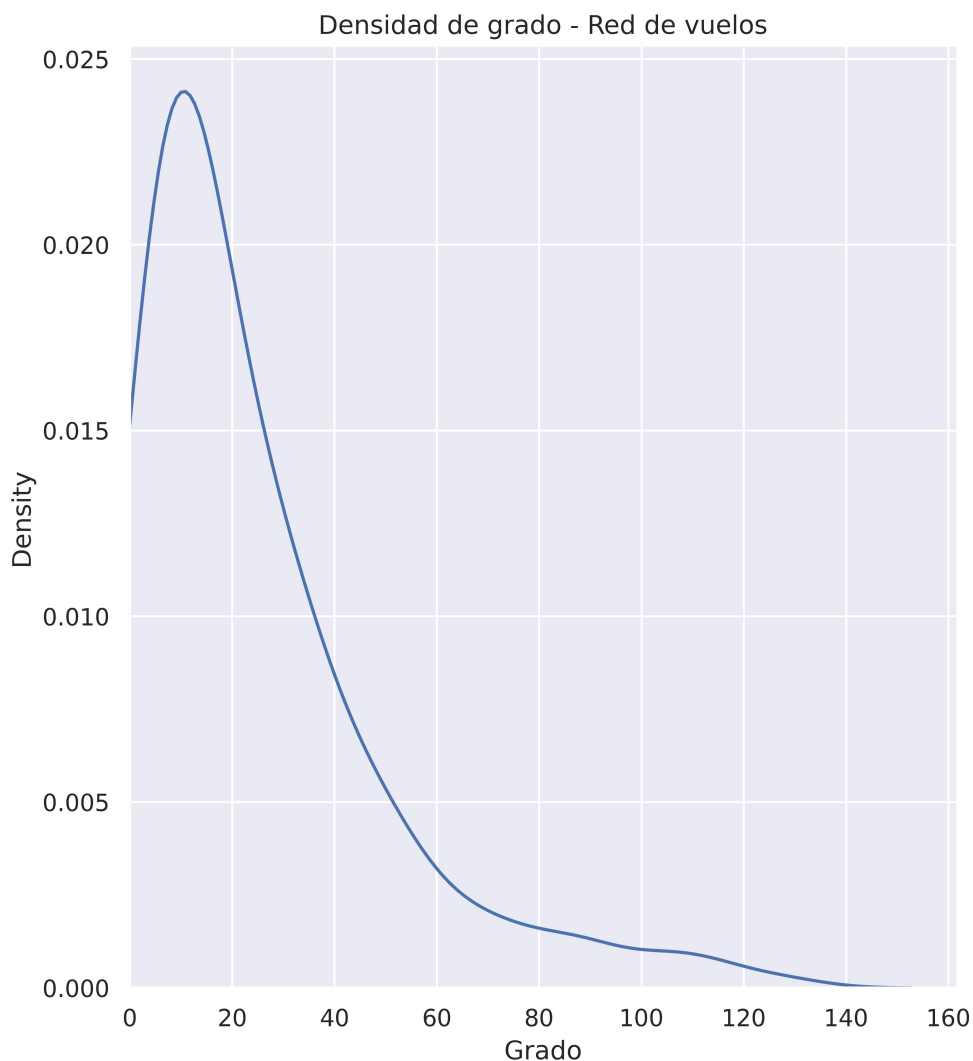


Se puede observar que Estados Unidos, Francia y el Reino Unido son los países con mayor puntaje de BC. Tiene sentido que aparezcan países grandes economicamente, ya que por lo son los más altos en tráfico aéreo, y por ende con una cantidad grande de rutas aéreas.

En el caso de Francia, la mayoría de los vuelos en los cuales es parte de un camino mínimo provienen de países origen/destino en África (lo cual me resulta extremadamente curioso...).

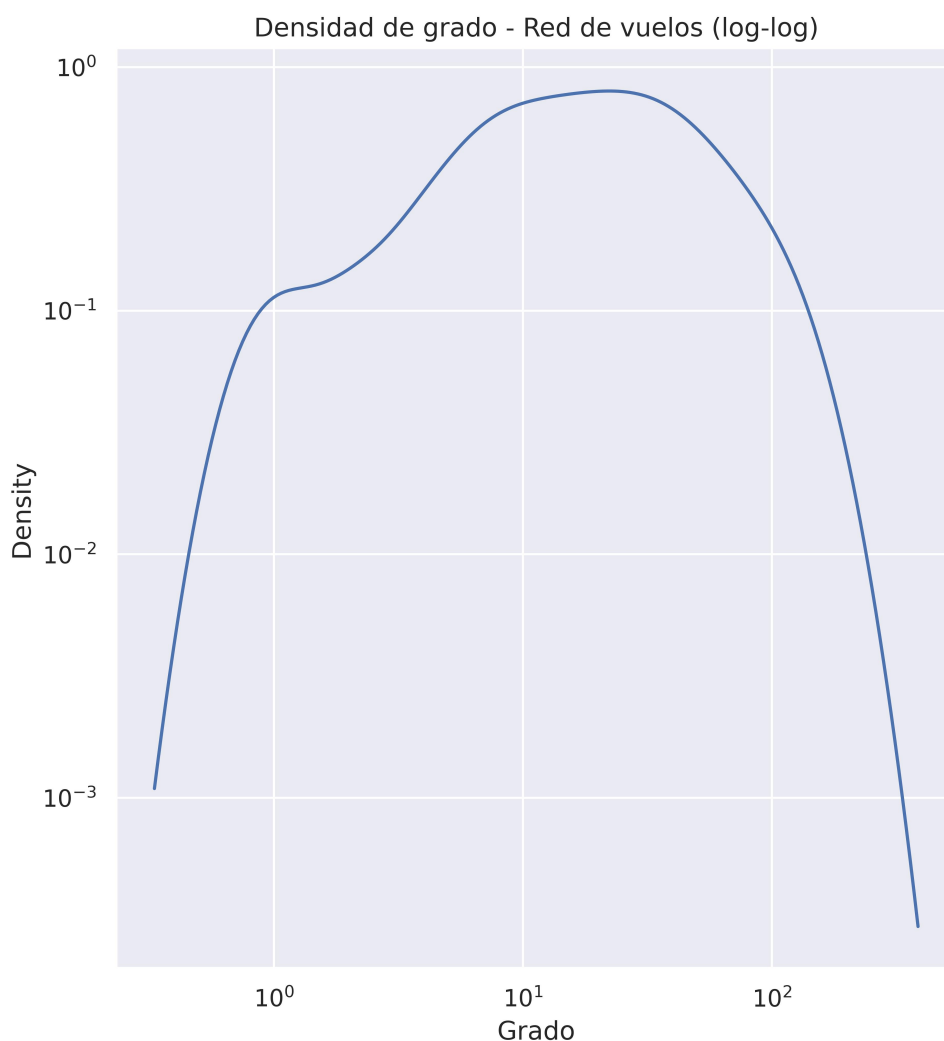
6. Modelados

En esta sección se muestran los modelos construidos a partir de los parámetros de la red. Se muestran en primer lugar las densidades de probabilidad calculadas para la red original.



Vemos la característica caída rápida de una power law en la densidad. El pico se encuentra aproximadamente en $x = 10$, denotando que la moda en cuanto al grado en la red es de 10.

En una escala log-log:



La caída en este caso no parece asemejarse completamente a una recta, pero se ve que empieza a darse entre $x = 10$ y $x = 20$.

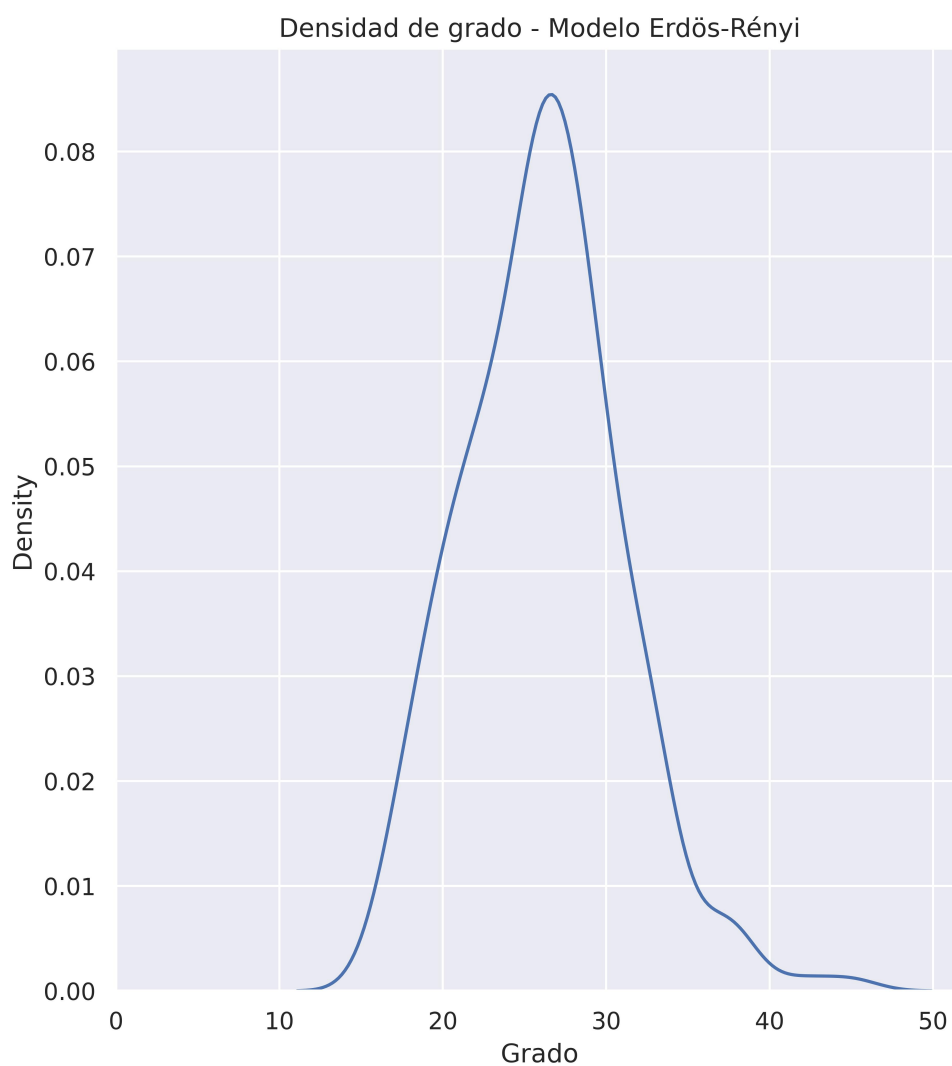
6.1. Erdős-Rényi

Para hacer un modelado de Erdős-Rényi de tipo $G(n, p)$, necesitamos obtener n , el número de vértices, y en este caso países, de la red, y p , la probabilidad que un par de países cualesquiera de la red esten conectados entre si. Utilizando la biblioteca de [modelos](#), en vez de proporcionarle p , le podemos proporcionar k , el grado promedio de la red. Sabemos que $p = k/(n - 1)$, por estimador insesgado del valor esperado. En la red original, calculamos $n = 229$ (países) y $k = 24,91$ ($p = 0,11$).

Muestro los resultados obtenidos:

Diámetro	Grado Promedio	C. de Clustering	Aristas
3	26.14	0.11	2993

Vemos que si bien la cantidad de aristas es parecida, el diámetro es chico y el grado promedio también es muy similar, el coeficiente de clustering es extremadamente bajo (solo 11 % de las aristas están en un triángulo). Cabe destacar que esta red también está llena de puentes locales, lo cual coincide con el bajo coeficiente de clustering.



Vemos que la densidad de grado tiene pinta de provenir de una distribución normal, como era de esperarse, ya que un modelado por Erdős-Rényi nunca podría generar una distribución que cumpla con una ley de potencias.

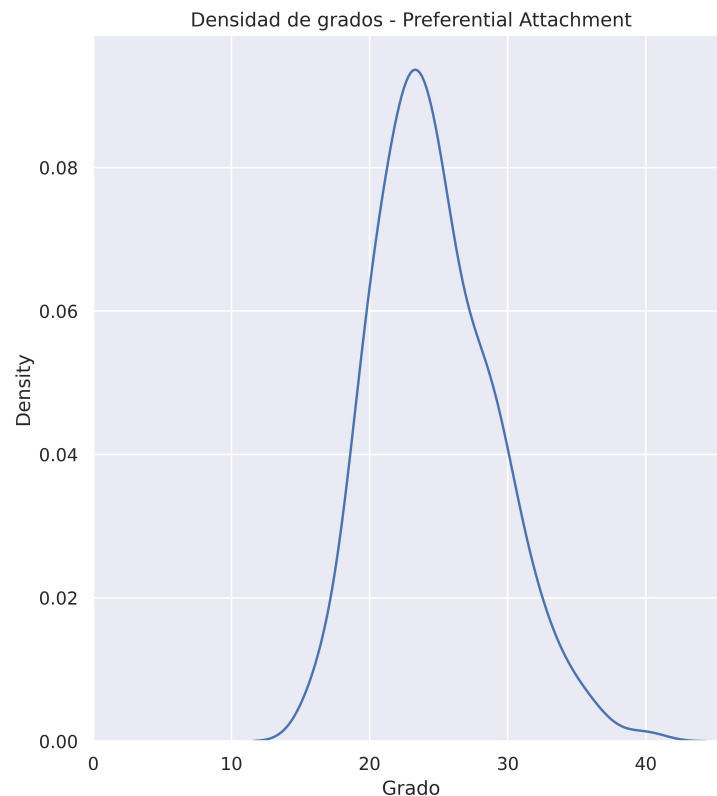
6.2. Preferential Attachment

Decidí realizar la estimación de alfa a partir del estimador de maxima verosimilitud para una ley de potencias. Lo que necesita este estimador es un valor de x a partir del cual se empieza a cumplir la ley de potencias. Por lo que se pudo ver del grafico de densidad, cuando $x = 10$, la densidad alcanza su pico y luego empieza a caer velozmente, por lo cual, luego de prueba y error, me termine decidiendo por $x_{min} = 10$.

Esto nos devuelve para la red original un valor de alpha estimado $\alpha = 2,45$, que cierra con los valores esperados para una red social. El otro hiperparámetro es k , la cantidad de aristas que se podrían llegar a agregar a la red al insertar un nuevo vértice. Para este parámetro, lo que hice fue ir probando distintos valores de k , hasta que encuentre aquel que resultaba en una red con una cantidad de aristas similares. Estime que $k = 10$ es la mejor aproximación sobre la cantidad de aristas, que dio el siguiente resultado:

Diámetro	Grado Promedio	C. de Clustering	Aristas
3	24.56	0.11	2812

Como en el modelo anterior, se puede ver que el coeficiente de clustering es extremadamente bajo. Si bien puede ajustarse un poco mejor a la distribución de grados, no es un modelo que pueda captar muy bien la forma de la red.



De hecho, la distribución calculada tiene una apariencia más similar a una distribución normal que una power law (aunque se puede notar una sutil asimetría positiva).

6.3. Anonymous Walks

Para comparar los dos modelos a la red original, obtenemos embeddings de la red original y los dos modelos mediante [Anonymous Walks](#).

Para realizar estos anonymous walks, tome como largo mínimo 2 y como largo máximo 8, ya que de 9 para arriba se consume demasiado tiempo y se necesitan recursos que no dispongo. Utilizando un thread pool de Python se pueden paralelizar los cálculos:

Largo A.W.	Distancia Coseno (Erdős-Rényi)	Distancia Coseno (Preferential Attachment)
2	0.00	0.00
3	6.78e-05	1.92e-05
4	1.00e-04	1.10e-04
5	1.90e-04	2.10e-04
6	3.40e-04	3.03e-04
7	5.03e-04	4.55e-04
8	6.24e-04	5.95e-04

Se puede apreciar que para caminos de longitud menor a 6, la distancia coseno es menor para el modelo de Erdős-Rényi, de manera tal que en ese caso se podría considerar más afín. Pero si consideramos que los embeddings de mayores dimensiones representan mejor a la red original, entonces el modelo de Preferential Attachment es el que mejor se ajusta, ya que la distancia coseno es menor para los caminos de longitud 6, 7 y 8 (y entiendo que esa tendencia seguiría si pudiese calcular los embeddings de mayores dimensiones).

Un dato curioso es que para los embeddings de tamaño 2, el cálculo de la distancia coseno es 0 para ambos resultados. Entiendo que esto tiene que ver con el hecho de que los graphlets de dimension 2 que se pueden levantar de las redes serán siempre parecidos (cualquiera sea la red) y que aparecen en las 3 redes de la misma manera.