Pre TP2: Data Mining en Ciencia y Tecnología

José Saint Germain
joesg998@gmail.com

1. Introducción

El análisis de la topología de grafos (es decir, redes) es un área de investigación que atañe a diferentes campos de estudio. Para ejemplificar el uso de grafos nos enfocaremos en el los datos obtenidos en el trabajo de Tagliazucchi y colaboradores (2013) que busca relacionar cambios en la modularidad de las redes construidas a partir de la señal de resonancia magnética funcional (fMRI) con los distintos estadíos del sueño.

2. Objetivos

Familiarizarse con la generación de grafos que representen un conjunto de datos. Visua- lizar, manipular y comparar distintos grafos. Calcular parámetros básicos de un grafo, y compararlos con modelos de redes random, small world y scale-free.

3. Estructura de los Datos

En la carpeta DataSujetos se encuentran los archivos separados por cada sujeto y estadio. Para cada sujeto y estadío de sueño encontraremos una matriz de correlaciones de tamaño 116x116 con las correlaciones entre las señales BOLD de 116 regiones cerebrales.

Además se incluyen los nombres y coordenadas de las 116 regiones en un archivo apar- te: aalextendedwithCoords.csv. Estas regiones están definidas a partir del atlas Automatic Anatomical Labeling (AAL). Ejemplos de los procedimiento para comenzar el análisis pueden encontrarse en este colab.

4. Preprocesamiento de los datos

Cargar el dataset con los datos para cada sujeto y los nombres y coordenadas de las regiones cerebrales a las que se les registró la actividad. Reportar cuántos sujetos y cuántos estados de sueño se observan en el conjunto de datos.

Cuadro 1: Conteo de estados de sueño en el conjunto de datos.

estado	cantidad de sujetos
N1	18
N2	18
N3	18
W	18

Como se observa en el cuadro 1, el dataset cuenta con 18 sujetos, teniendo cada uno lo cuatro diferentes estados de sueño.

5. Manipulación de datos

5.1 Graficar la matriz de correlaciones entre regiones (es decir, la "matriz de adyacencia pesada") para el sujeto 2 de la condición despierto ("Wake"). Transformar dicha matriz de adyacencia pesada a una matriz de adyancia binaria $A_{i,j}$ que represente una una densidad de enlaces igual a 0.08. ¿Cuál es el valor de umbral de correlación entre pares de regiones que tuvo que utilizar?

Como se observa en la figura 1, se grafica en la imagen de la izquierda la matriz de correlaciones entre regiones. A su vez, en la imagen de la derecha, se grafica la misma matriz representando una densidad de enlaces igual a 0.08. Para ello, se utilizó un umbral de 0.75, el cual se obtiene con la función desarrollada por el profesor, utilizando como hiperparámetro la densidad de enlaces deseada.

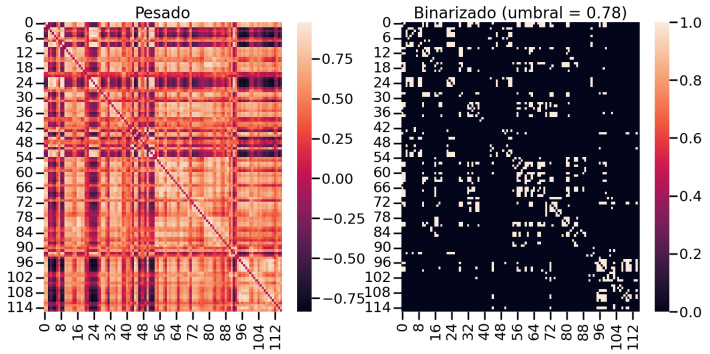

Figura 1: Matriz de adyacencia pesada y binarizada del sujeto 2 despierto

5.2 Utilizando $A_{i,j}$, obtener el grafo resultante G ¿Es G un grafo conectado? ¿Se puede calcular la distancia media entre pares de nodos d del grafo G? ¿Si no se puede, qué medida equivalente calcularías?

El grafo que se obtiene de $A_{i,j}$ no está conectado, por lo que no se puede calcular la distancia media entre pares de nodos. Alternativamente, se puede calcular la distancia media de su componente gigante (es decir, el componente conectado más grande dentro del grafo G), el cual es 3.85. Otra opción es calcular la eficiencia global del grafo el cual se calcula en el ejercicio siguiente.

5.3 Calcular d para cada componente conectado de G. Calcular la eficiencia global ef f del grafo G.

La distancia d de los dos componentes conectados más grandes de G dan 3.85 y 1.28. Después, el resto de componentes conectados tienen un valor de d igual a 0. Por último, la eficiencia global del grafo G es 0.2446.

5.4 Obtener la lista de enlaces del grafo G. Calcular el grado promedio <k >, el nodo con grado máximo kmax, el coeficiente de clustering promedio

Cuadro 2: Enlaces aleatorios del grafo G

Nodo 1	Nodo 2
23	92
10	44
11	62
28	99
4	53
60	66
0	28
95	97

Cuadro 3: Medidas del grafo G

Medida	Valor
Grado promedio (K)	9.207
Nodo con grado máximo (kmax)	30
Coeficiente de clustering promedio (C)	0.527
Eficiencia global	0.245

Ya que la lista de enlaces del grafo es extensa. Se muestra una lista aleatoria de 8 enlaces del grafo G. También se muestran los valores de las medidas del grafo G en el cuadro 3.

5.6 Visualizar el grafo, ubicando los nodos en sus coordenadas cerebrales y coloreando cada nodo de acuerdo a su coeficiente de clustering Ci. Graficar la distribución de grado del grafo, elijiendo un número de bins apropriado

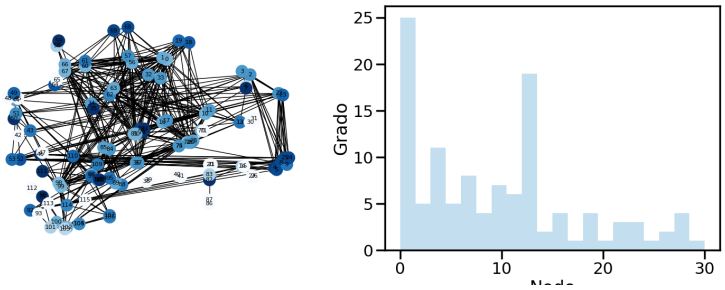

Figura 2: Grafo G

5.7 Vamos a comparar el grafo G con prototipos de redes poissonianas (random), small-World y scale-free, usando los algoritmos de Erdos-Renyi, Watts-Strogatz y Barabasi-Albert, respectivamente. Para ello, elegir (y reportar) los parámetros utilizados para cada algoritmo, buscando siempre que los grafos simulados de dichos prototipos sean comparables al grafo de datos G (en términos de número de nodos y números de enlaces). Visualizar un ejemplo de grafo para cada uno de estos prototipos de redes. Discutir diferencias.

A continuación, se enlistan los distintos parámetros utilizados para generar cada tipo de grafo:

Cuadro 4: Parámetros de las redes aleatorias

Red	Nodos	Enlaces	Parámetro 2 (scale free y small world)		Parámetro 3 (small world)	
Poisson	116	534	-	-	-	-
Small world	116	464	Vecinos conectados a cada nodo	9	Prob. de re-conexión	0.02
Libre de escala	116	555	Q de enlaces nuevos por nodo nuevo	5	-	-

Nota: para la red poissoniana no se aclaran parámetros extra porque los que se utilizaron fueron al cantidad de nodos y la de enlaces.

Adicionalmente, se grafican los mismos con una disposición basada en el algoritmo de Fruchterman Reingold:

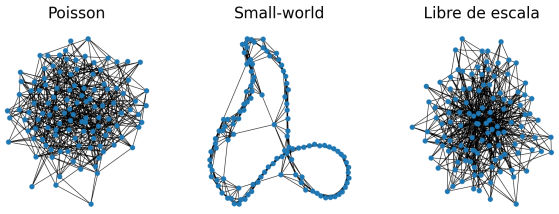

Figura 3: Visualización de un grafo de cada prototipo

Con respecto a la red poissoniana, no hay mucho para destacar, dado su comopnente aleatorio. La red Small-world llama la atención por tener muchos nodos encadenados, con varios grupos de nodos sin esa forma. Por último, la red libre de escala a simple vista tienen una forma muy similar a la poissoniana.

5. Generar 1000 instancias de grafos para cada uno de dichos prototipos (poissonianas, small-World y scale-free). Para el conjunto de 1000 instancias de cada prototipo, calcular el histograma de coeficientes de <k>, kmax, C, y eff. Comparar con los valores de coeficientes que obtuvimos para el grafo de datos G.

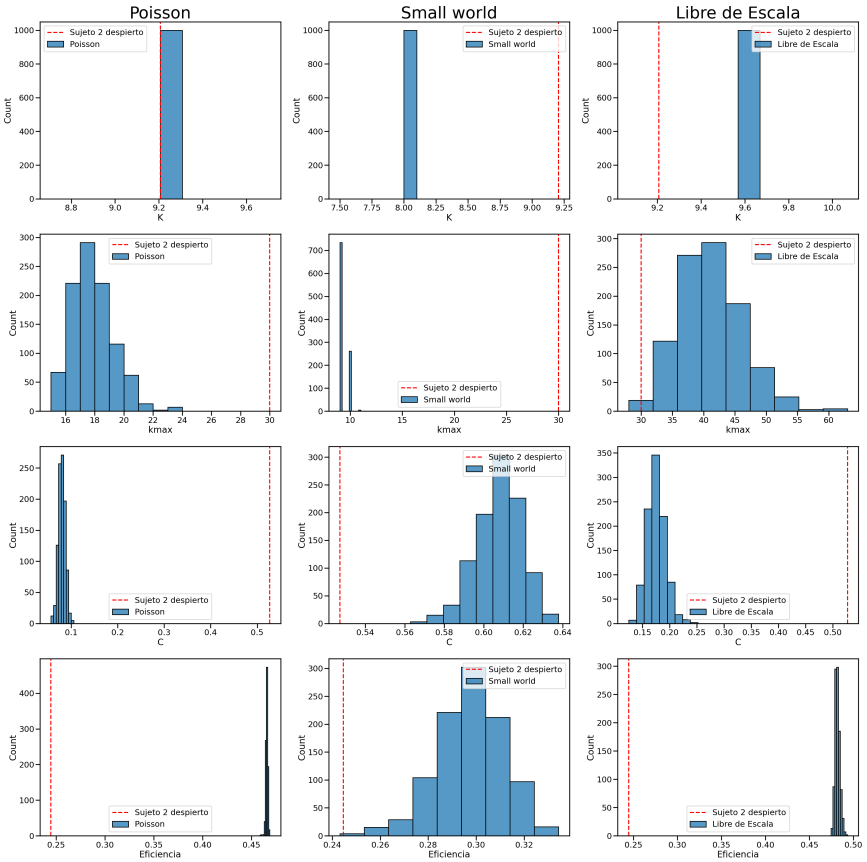

Figura 4: Comparación de métricas de grafo G con el resto

Con respecto al grado promedio del grafo (K), se observa que el grafo G tienen el mismo valor que la red poissoniana. La red Small-world tiene un menor grado promedio, lo cual hace sentido con la imagen observada anteriormente, en donde una gran cantidad de nodos apenas se conectan con otros dos. Por último, la red libre de escala tiene un grado promedio mayor que el grafo G; indicando la presencia de hubs con alto grado.

En cuanto al grado máximo (kmax), se observa que el grafo G tiene un valor mayor que la Poisson. También es mayor que el Small-world, cuyos valores están más concentrados que en la Poisson. Por último, el grado máximo de G aparece entre los valores más chicos del grado máximo de la red libre de escala, acompañando la misma idea planteada para el grado promedio.

Tomando el coeficiente de clustering medio (C), el grafo G tiene un valor mayor que la red Poisson, indicando que los vecinos de cada nodo están mucho más conectados. Por otro lado, el Small-world tiene un valor más alto. Se puede entender que al tener una probalidad de reconexión baja (0.02) la

mayoría de los nodos están conectados con su primer y segundo vecino, los cuales estarían conectados entre sí. Después, el libre de escalas tiene valores menores al del G, indicando la presencia de hubs, que conectan nodos que no están conectados entre sí.

Finalmente, la efiencia global (eff) del grafo G es mucho menor que las tres redes. Esto evidencia que la distancia promedio de los nodos de G es mucho mayor. De todas formas, el valor de G se acerca a los valores mínimos de la Small-world, lo cual puede entenderse si tiene una probabildiad de reconexión baja ya que es muy probable que tengas pocos atajos (enlace recableado") para llegar de un nodo a otro.