
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 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 estadios del sueño.

2. Objetivos

Familiarizarse con la generación de grafos que representen un conjunto de datos. Visualizar, 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 estadio 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 aparte: aalexten- dedwithCoords.csv. Estas regiones están definidas a partir del atlas Automatic Anatomical Labeling (AAL). Ejemplos de los procedimientos 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 los cuatro diferentes estados de sueño.

24 5. Manipulación de datos

25 5.1 Graficar la matriz de correlaciones entre regiones (es decir, la "matriz de adyacencia pesada")
 26 para el sujeto 2 de la condición despierto ("Wake"). Transformar dicha matriz de adyacencia pesada
 27 a una matriz de adyacencia binaria $A_{i,j}$ que represente una densidad de enlaces igual a 0.08.
 28 ¿Cuál es el valor de umbral de correlación entre pares de regiones que tuvo que utilizar?

29 Como se observa en la figura 1, se grafica en la imagen de la izquierda la matriz de correlaciones
 30 entre regiones. A su vez, en la imagen de la derecha, se grafica la misma matriz representando una
 31 densidad de enlaces igual a 0.08. Para ello, se utilizó un umbral de 0.75.

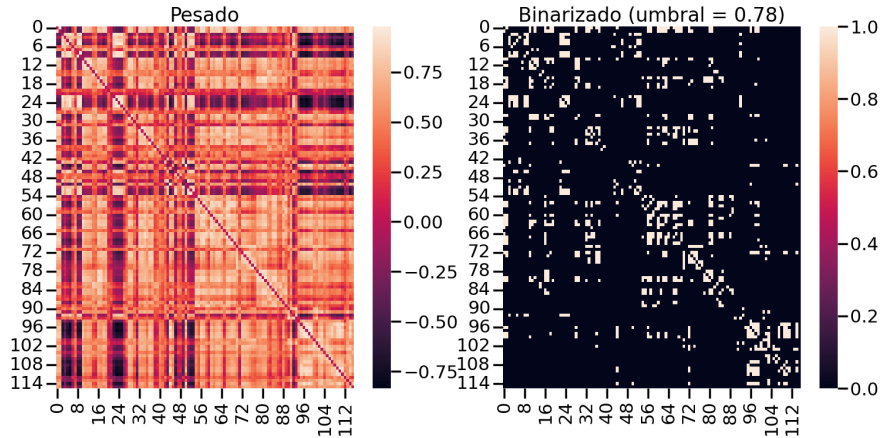


Figura 1: Matriz de adyacencia pesada y binarizada del sujeto 2 despierto

32 5.2 Utilizando $A_{i,j}$, obtener el grafo resultante G . ¿Es G un grafo conectado? ¿Se puede calcular
 33 la distancia media entre pares de nodos d del grafo G ? ¿Si no se puede, qué medida equivalente
 34 calcularías?

35 El grafo que se obtiene de $A_{i,j}$ no está conectado, por lo que no se puede calcular la distancia media
 36 entre pares de nodos. Alternativamente, calculamos la distancia media de su componente gigante (es
 37 decir, el componente conectado más grande dentro del grafo G), el cual es 3.85.

38 5.3 Calcular d para cada componente conectado de G . Calcular la eficiencia global eff del grafo G .

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

42 5.4 Obtener la lista de enlaces del grafo G .

43 Ya que la lista de enlaces del grafo es extensa. Se una lista al azar de 8 enlaces del grafo G :

Cuadro 2: Primeros 8 enlaces del grafo G .

Nodo 1	Nodo 2
28	96
109	110
73	75
72	99
16	57
44	110
8	22
19	81

- 44 5.5 Calcular el grado promedio $\langle k \rangle$, el nodo con grado máximo k_{max} , el coeficiente de clustering
 45 promedio
 46 A continuación se muestran los valores de las medidas del grafo G en el cuadro 3:

Cuadro 3: Medidas del grafo G

Medida	Valor
Grado promedio (K)	9.206897
Nodo con grado máximo (kmax)	30.000000
Coeficiente de clustering promedio (C)	0.527085
Eficiencia	0.244630

- 47 5.6 Visualizar el grafo, ubicando los nodos en sus coordenadas cerebrales y coloreando cada nodo
 48 de acuerdo a su coeficiente de clustering C_i

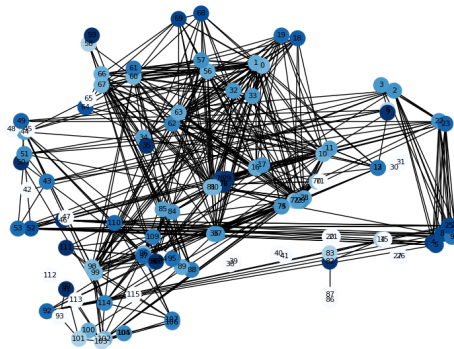


Figura 2: Grafo G

- 49 5.7 Graficar la distribución de grado del grafo, eligiendo un número de bins apropiado

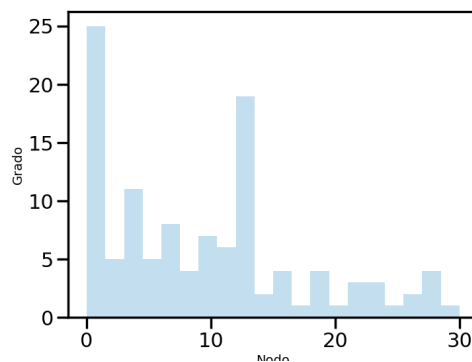


Figura 3: Distribución de grado del grafo

- 50 5.8 Vamos a comparar el grafo G con prototipos de redes poissonianas (random), small-World y
 51 scale-free, usando los algoritmos de Erdos-Renyi, Watts-Strogatz y Barabasi-Albert, respectivamente.
 52 Para ello, elegir (y reportar) los parámetros utilizados para cada algoritmo, buscando siempre que
 53 los grafos simulados de dichos prototipos sean comparables al grafo de datos G (en términos de
 54 número de nodos y números de enlaces). Visualizar un ejemplo de grafo para cada uno de estos
 55 prototipos de redes. Discutir diferencias.

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

Cuadro 4: Parámetros de las redes aleatorias

Red	Nodos	Parámetro 2		Parámetro 3	
Poissoniana	116	Prob. creación de enlaces	0.08	-	-
Small-world	116	Prob. re-conexión	0.20	-	-
Libre de escala	116	N de enlaces nuevos por nodo nuevo	4	Prob. conexión inicial	0.081

57 Adicionalmente, se grafican los mismos:

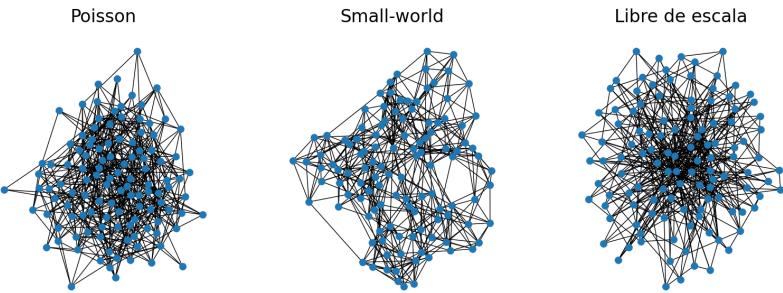


Figura 4: Visualización de un grafo de cada prototipo

58 5.9 Generar 1000 instancias de grafos para cada uno de dichos prototipos (poissonianas, small-
59 World y scale-free). Para el conjunto de 1000 instancias de cada prototipo, calcular el histograma de
60 coeficientes de $\langle k \rangle$, k_{max} , C , y eff . Comparar con los valores de coeficientes que obtuvimos para
61 el grafo de datos G .

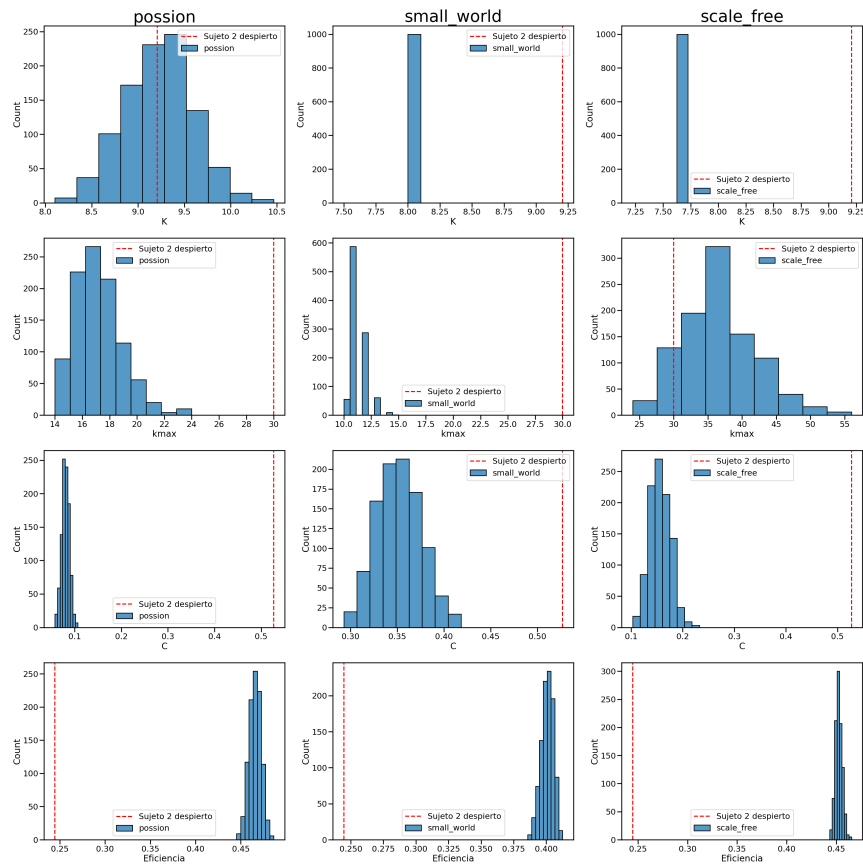


Figura 5: Comparación de métricas de grafo G con el resto