

Proyecto 1. Estructura en redes complejas: índices de centralidad.

Miruna Andreea Gheata

Profesor: Gabriel Cardona Juanals

Primera práctica de la asignatura de Redes Complejas del Máster Universitario en Sistemas Inteligentes (MUSI)

Universitat de les Illes Balears

07122 Palma, Islas Baleares, España

miruna.gheata1@estudiant.uib.cat

Resumen—Análisis de una red compleja obtenida a partir de datos recogidos mediante *web scraping*. Los datos empleados han sido diferentes enfermedades y sus síntomas. A partir de la red creada se han analizado sus características y aplicado diferentes índices de centralidad para determinar qué nodos son más importantes.

Index Terms—redes complejas, índices de centralidad, eigenvector, closeness, betweenness

I. INTRODUCCIÓN

En este documento se presenta el trabajo realizado para la primera práctica de la asignatura de Redes Complejas del Máster Universitario en Sistemas Inteligentes. A continuación se detalla el contenido de cada apartado.

En el Capítulo 2 se define la red compleja elegida. Se explicará la fuente de los datos que componen la red, su extracción y posterior transformación a una red.

En el Capítulo 3 se realiza un análisis de la red compleja resultante, enumerando las características que tiene, identificando los grupos de nodos que la componen y el modelo de red que sigue.

Por último, en el Capítulo 4 se aplican diferentes índices de centralidad a la red y se comparan los resultados obtenidos.

II. LA RED COMPLEJA DE SÍNTOMAS

Parte del trabajo que se ha realizado para esta práctica se ha dedicado a la *elección y recopilación de los datos* que compondrán la red compleja a analizar.

Los datos que se han elegido para crear la red compleja son las **enfermedades y los diferentes síntomas que las componen**. El tema puede resultar muy interesante debido a que al estar en forma de grafo, se obtiene una representación visual donde se puede observar qué síntomas se relacionan entre sí con más frecuencia. Además, también se puede llegar a determinar qué síntomas son los más comunes en las diferentes enfermedades que existen.

II-A. Obtención de los datos

Los datos se han recopilado de la página web de Medicinet [1]. En [1] se presenta un índice de síntomas organizados por orden alfabético. Cada síntoma tiene un enlace web equivalente donde se enumeran las diferentes enfermedades que contienen dicho síntoma. En la Figura 1 se muestran diferentes enfermedades relacionadas con el síntoma *Fever*.

Para poder recoger estos datos se ha empleado la técnica de *web scraping* que, tal como se explica en [2], “consiste en navegar automáticamente una web y extraer de ella información“. Existen diferentes programas de software y librerías encargadas para realizar este proceso. En esta práctica se ha empleado BeautifulSoup [3], una librería de Python que sirve para obtener información de elementos HTML y XML.

Los pasos para la obtención de los datos son:

1. Acceder a todos los síntomas [1].
2. Para cada síntoma:
 - a) Acceder al enlace que contiene su información.
 - b) Identificar las diferentes enfermedades que presentan dicho síntoma.
 - c) Guardar en un diccionario la información obtenida de manera que las llaves del diccionario sean las diferentes enfermedades y los valores del diccionario sean la lista de los diferentes síntomas que contiene cada enfermedad. Por ejemplo, el siguiente abstracto se corresponde con algunas entradas del diccionario de enfermedades tras analizar el síntoma *Fever*:

```
'Juvenile_Rheumatoid_Arthritis': ['Fever']
'Balamuthia': ['Fever']
'Bile_Duct_Cancer': ['Fever']
'Mantle_Cell_Lymphoma': ['Fever']
```

En el caso que una enfermedad ya pertenezca al diccionario, se accede a su lista de síntomas correspondiente y se añade el síntoma que se esté analizando.

3. Guardar el diccionario resultante dentro de un fichero llamado *diseases.txt*.

El script empleado para realizar el *scraping* se encuentra en este fichero del repositorio de Git [4].

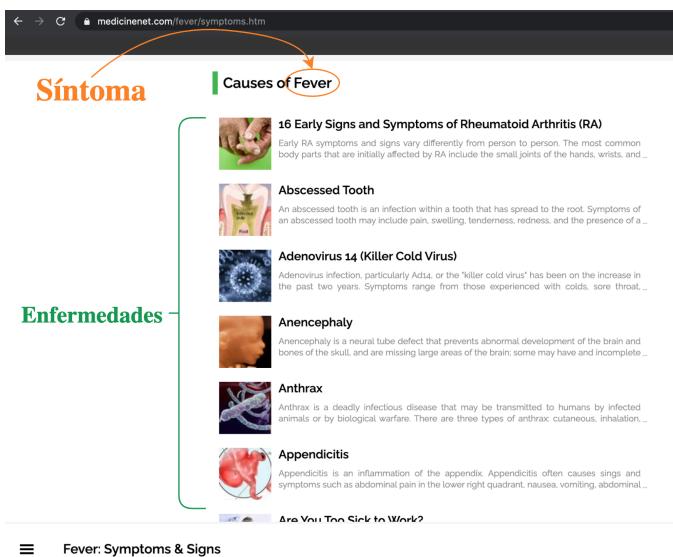


Figura 1: Algunas de las enfermedades que están relacionadas con el síntoma *Fever* (Fuente: [5]).

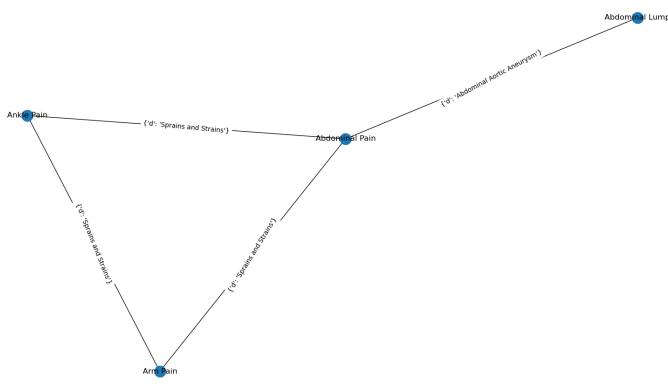


Figura 2: Ejemplo de subgrafo de la red compleja de síntomas.

II-B. Red compleja resultante

Tras la obtención de los datos en bruto (es decir, el fichero diseases.txt) se ha creado la propia red compleja. Una red está compuesta por nodos y por aristas que unen dichos nodos. Para este caso, se ha elegido que:

- Los nodos son los diferentes *síntomas* que puede presentar una persona.
 - Las aristas que unen los síntomas son las *enfermedades* que comparten los síntomas presentados.

Como ejemplo, en la Figura 2 se ilustra un grafo que se ha formado a partir de estos datos:

1. Enfermedad *Cyst* con síntoma Abdominal Lump.
 2. Enfermedad *Sprains and Strains* con síntomas Arm pain, Abdominal Pain, Ankle Pain.
 3. Enfermedad *Abdominal Aortic Aneurysm* con síntomas Abdominal Pain, Abdominal Lump.

Nodos	1501
Aristas	21708
Grado medio	28.87
Conexo	Falso
Componentes conexas	42

Cuadro I: Características del grafo original.

Debido al elevado número de datos recogidos dentro de diseases.txt, se ha optado por emplear un subgrupo de estos datos para la creación de la red. En concreto, este subgrupo se caracteriza por ser compuesto por enfermedades que tienen entre **2** y **50** síntomas.

III. ANÁLISIS DE LA RED COMPLEJA

III-A. Definiciones

Sea G el grafo que se está analizando, V el número de nodos n de G y E el número de aristas m del mismo. Formalmente, el grafo se define como:

$$G = V, E \quad (1)$$

Este grafo es un grafo **no dirigido**, ya que no es necesario representar el sentido de la dirección para esta red.

Sea c el grado medio del grafo:

$$c = \frac{1}{n} \sum_i k_i = \frac{2m}{n} \quad (2)$$

Sea D la densidad del grafo:

$$D = \frac{2|E|}{|V|(|V|-1)} \quad (3)$$

Sea T la transitividad del grafo:

$$T = \frac{3 \times \text{número de triángulos del grafo}}{\text{total de nodos triples conectados en el grafo}} \quad (4)$$

III-B. Características de la red

En el Cuadro I se muestran las características del grafo resultante. Como se puede observar, existen 1501 nodos diferentes (es decir, síntomas) y 21708 aristas diferentes (que se corresponden con las enfermedades). Cabe destacar que **las aristas se han reducido** de manera que si dos síntomas tienen más de una enfermedad en común, la arista que los une tendrá un peso equivalente al total de enfermedades que comparten. En la Figura 3 se ilustra la red original.

Este primer grafo creado no es conexo; existen 42 componentes diferentes que no están conectadas. Para el análisis que se desea realizar, cogeremos la componente conexa más grande que presenta este grafo original.

La información de la componente elegida se encuentra en el Cuadro II. Analicemos detenidamente cada elemento:

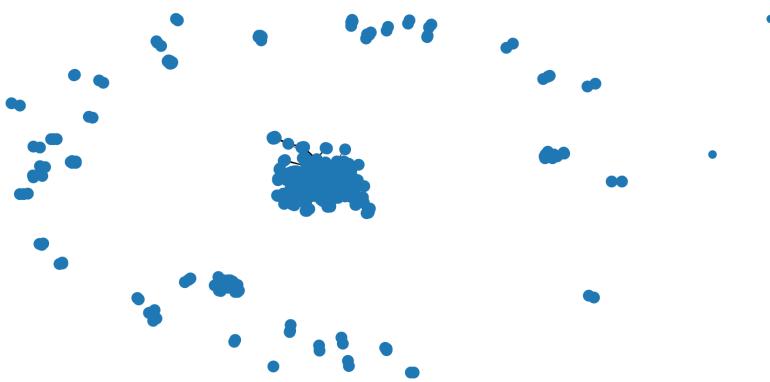


Figura 3: Grafo obtenido a partir de los datos de diseases.txt.

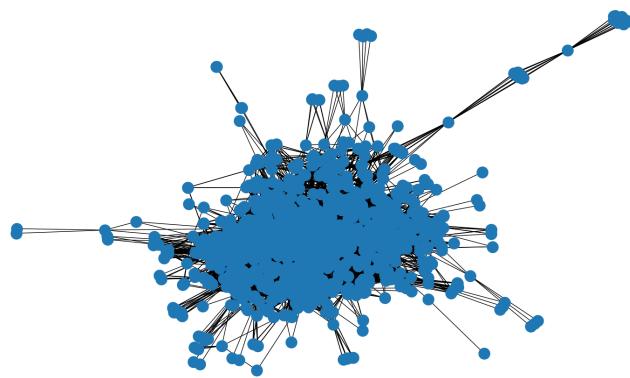


Figura 4: Subgrafo equivalente a la componente más grande del grafo original (Figura 3)

1. El **total de nodos** de este grafo es de **1345**. Es decir, existen 1345 síntomas diferentes.
2. El **total de aristas** es de **21396**. Es decir, existen al menos 21396 enfermedades diferentes (recordemos que las aristas tienen como peso el total de enfermedades que unen dos síntomas).
3. El **grado medio** de los nodos (Ecuación 2) es de **31.81**. Esto quiere decir que:
 - El número medio de enfermedades que presentan un síntoma concreto es 31 aproximadamente.
 - Para cada síntoma de media está relacionado con otros 31 síntomas distintos.

Nodos	1345
Aristas	21396
Grado medio	31.81
Conexo	Verdadero
Diámetro	8
Distancia media	3.0789
Densidad	0.0236
Transitividad	0.6219
Coeficiente clustering	0.765
Puntos de articulación	14
Puentes	5

Cuadro II: Características del grafo analizado.

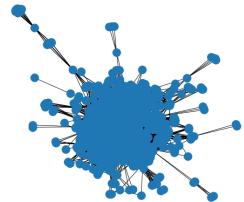


Figura 5: Red resultante tras eliminar el punto de articulación con mayor grado.

4. El **diámetro** del grafo es **8**. Esto implica que el camino más largo entre los 2 nodos que más alejados estén el uno del otro es de 8.
5. La **distancia media** entre los nodos es de **3.0789**. Por lo tanto, los nodos están muy cerca el uno del otro.
6. La **densidad** del grafo (Ecuación 3) es de **0.0236**. La densidad es el ratio entre las aristas del grafo y el total de aristas que podría tener del grafo. Esto indica que los nodos del grafo no están muy conectados, es decir, es *poco denso*.
7. La **transitividad** del grafo es de **0.6219**. Indica la probabilidad de que los nodos adyacentes estén interconectados [6]. Esto muestra que hay unas comunidades formadas entre los diferentes nodos del grafo.

Por otra parte, tenemos que la red tiene **14 puntos de articulación** y **5 puentes**. Ambos elementos sirven para mantener conexa la red, ya que la eliminación de un punto de articulación o de un puente implicaría que una o más componentes previamente conexas quedarían inaccesibles.

En las Figuras 5 y 6 se muestra el grafo en el caso de que se elimine el nodo con mayor grado o uno de los puentes, respectivamente. En el primer caso la nueva red estaría compuesta por 3 componentes, y en el segundo estaría compuesta por 2 componentes.

Otra medida interesante es el **coeficiente de clustering**. En la Figura 7 se muestra el histograma de los coeficientes de *clustering* de la red. La mayoría de nodos tienen un alto coeficiente de *clustering*, por lo que significa que hay muchos síntomas que están muy agrupados entre si. EL coeficiente de *clustering* medio de la red es 0.765.

III-C. Modelo de la red

Las redes se pueden clasificar según su distribución de grados en distintos tipos de modelos de red. Sin embargo, para la red seleccionada no se aprecia qué tipo de modelo es

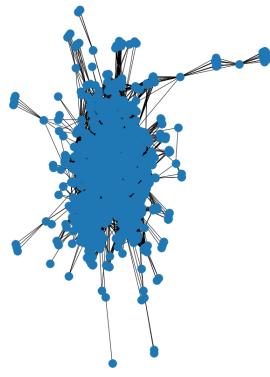


Figura 6: Red resultante tras eliminar uno de los puentes de la red.

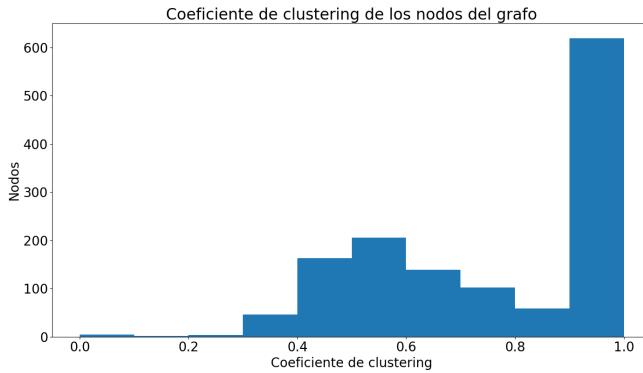


Figura 7: Coeficientes de *clustering* de los nodos de la red.

(Figura 16). Esto posiblemente se deba a que se ha realizado la selección inicial de los nodos (enfermedades con síntomas entre 2 y 50) de manera directa, y no de forma aleatoria.

III-D. Grupos de nodos

Dentro de la red existe se pueden determinar grupos cohesionados de nodos. En los siguientes apartados se determinarán el clique, los cores y las comunidades de la red.

III-D1. Clique: Un clique es un subgrafo máximo *completo* de la red; un grafo completo es aquel en el que todos los nodos están conectados al resto de nodos, es decir, tienen el mismo grado. En la Figura 8 se muestra el clique de la red. En este caso, el clique está compuesto por 42 nodos, 861 aristas y tiene un grado medio de 41.

III-D2. Cores: Tal como se define en [7], “un k-core es un subgrafo máximo que contiene nodos de grado k o superior”. En las Figuras 10, 11 y 12 se ilustran los nodos con grado $k \geq 30$, $k \geq 40$ y $k = 48$, respectivamente. El k-core máximo de esta red es con $k = 48$.

En la Figura 9 se muestran todos los nodos coloreados según el k-core al que pertenecen. Cuanto más oscuro sea el color del nodo, mayor será su grado.

III-D3. Comunidades: Tal como se menciona en [8], “una comunidad consiste en un grupo de nodos que están altamente conectados entre ellos pero escasamente conectados a otros grupos densos de la red”.

Ésta red tiene 10 comunidades diferentes y se muestran en la Figura 13. La comunidad más grande contiene 388 nodos, y las más pequeñas contienen sólo 4 nodos. En las Figuras 14 y 15 se descomponen las 10 comunidades diferentes de la red, mostrando para cada una el total de nodos que pertenecen a cada una de ellas.

IV. ÍNDICES DE CENTRALIDAD

Los índices de centralidad sirven para poder determinar la importancia de cada nodo dentro de la red. De todos los índices diferentes de centralidad que existen se han analizado 4: la *centralidad de grados*, el *eigenvector*, el *closeness* y el *betweenness*.

A continuación se enumeran los diferentes síntomas que han aparecido como más importantes según los diferentes índices de centralidad:

1. Bad breath, Enlarged glands and Pain
2. Bad taste in mouth, Difficulty swallowing, Enlarged glands and Sore throat
3. Bleeding gums, Ear ache, Enlarged or swollen glands and Pain when moving eyes
4. Bloating, Diarrhea and Stomach cramps
5. Bloating, Nausea or vomiting, Stomach cramps and Upset stomach
6. Bloody Sputum
7. Blurred Vision
8. Blurred vision, Cloudy vision, Enlarged or swollen glands and Fear of air
9. Body aches or pains, Fatigue, Tires quickly and Weight loss (unintentional)
10. Brittle hair, Change in hair texture, Coarse hair and Dry skin
11. Brittle hair, Change in hair texture, Dry skin and Fatigue
12. Brittle hair, Change in hair texture, Hair loss and Pain or discomfort
13. Brittle hair, Dry skin (General), Dry skin (Skin) and Fatigue
14. Bruising or discoloration, Drooping eyelid, Eye irritation and Eyelid redness
15. Bulging neck veins, Enlarged glands and Pain
16. Bumps on Skin

IV-A. Centralidad de grados

La centralidad de grados sirve para determinar el número de enlaces que tiene cada nodo. Por lo tanto, los nodos más importantes serán aquellos que tengan **el grado más elevado**.

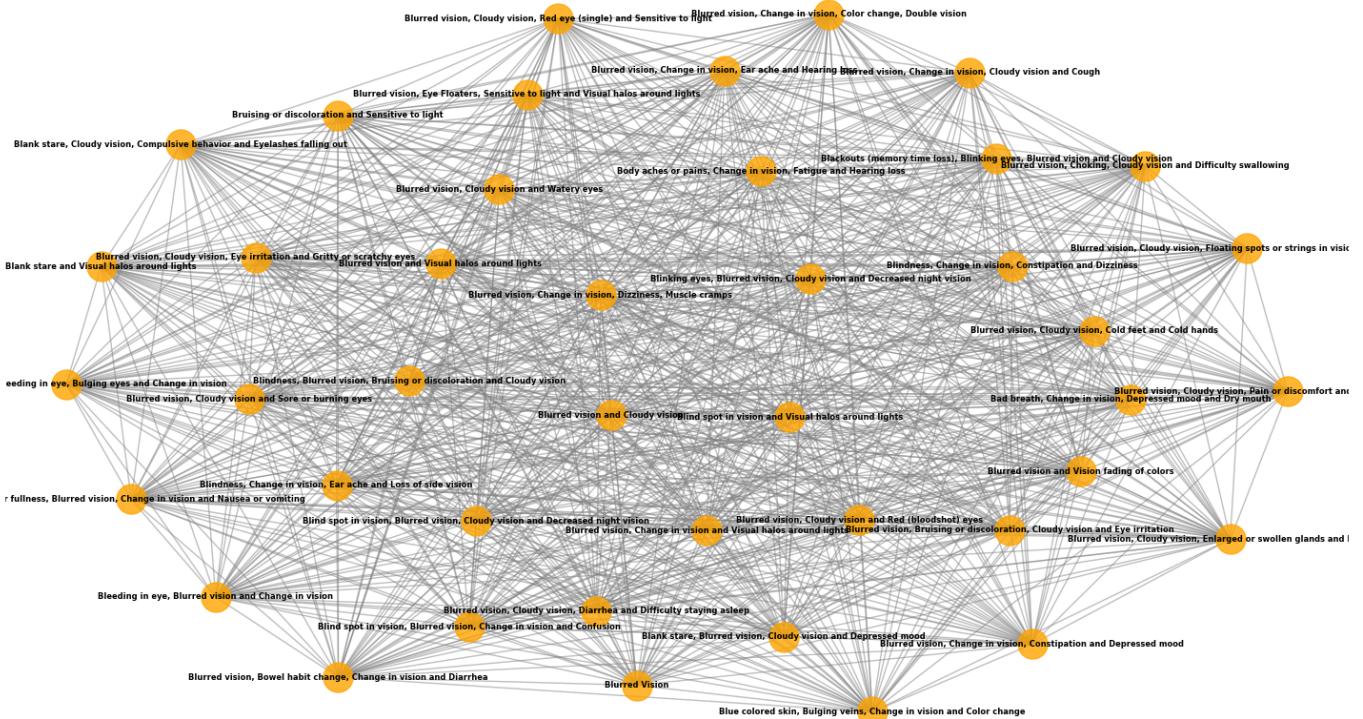


Figura 8: Clique de la red.

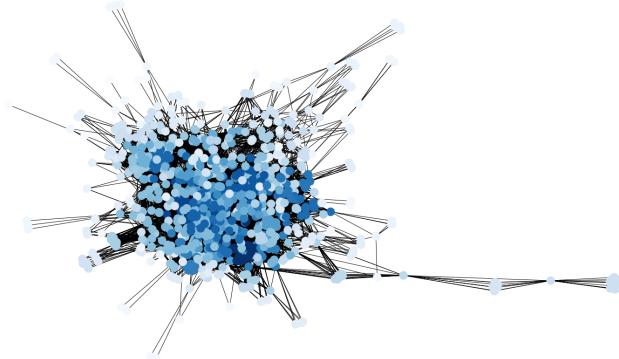


Figura 9: Nodos coloreados según el core al que pertenecen. Cuanta mayor grado tenga un nodo más oscuro es su color.

En la Figura 16 se muestra el histograma de los grados de todos los nodos de la red. Como se puede observar, el grado de los nodos está entre un rango de 2 y 52, (recordemos que vimos que esta red tiene un grado medio de 31.81).

En la Figura 17 se muestra un histograma con los valores obtenidos con la medida de grados de centralidad. La mayoría de nodos de la red tienen ha obtenido grado menor a 0.04.

En el Cuadro III se muestran los 5 síntomas que han obtenido un mayor valor para esta medida, y se tiene que el síntoma de ***Blurry Vision*** es el más importante.

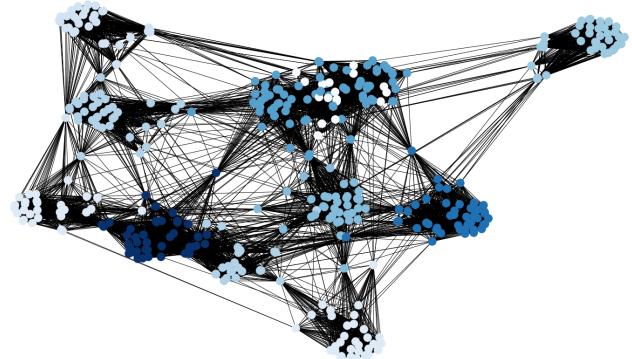


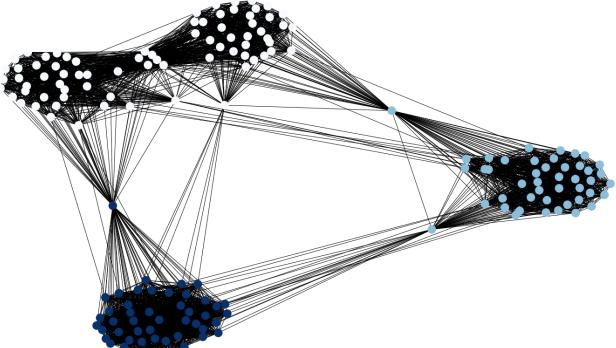
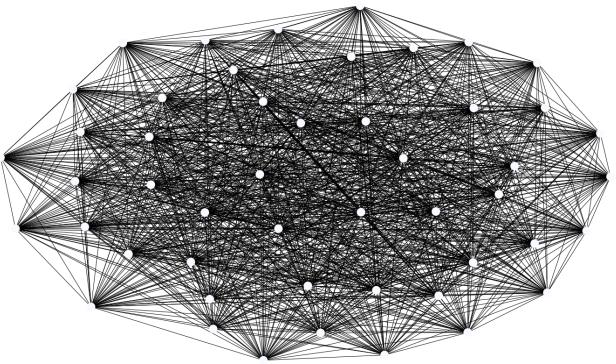
Figura 10: Nodos core $k \geq 30$.

Posición	Síntoma	Centralidad de Grado
1	7	0.09375
2	10	0.07589
3	12	0.07514
4	14	0.07291
5	15	0.07142

Cuadro III: Los 5 síntomas más importantes según la centralidad de grado.

IV-B. Eigenvector

La medida de *eigenvector* analiza la influencia de un nodo dentro de la red. Como bien se explica en [9], “una persona con pocas conexiones podría tener un mayor coeficiente de

Figura 11: Nodos core $k \geq 40$.Figura 12: Nodos core $k = 48$.

eigenvector si esas pocas conexiones le une a otros que están muy conectados. Esta centralidad permite que las conexiones tengan un valor variable, de manera que conectarse a unos nodos concretos tiene mayor beneficio que conectarse a otros.“

En la Figura 18 se muestra el histograma de coeficientes de *eigenvector* de todos los nodos de la red. La mayoría de nodos tienen un valor inferior al 0.04.

En el Cuadro IV se muestran los 5 síntomas que han

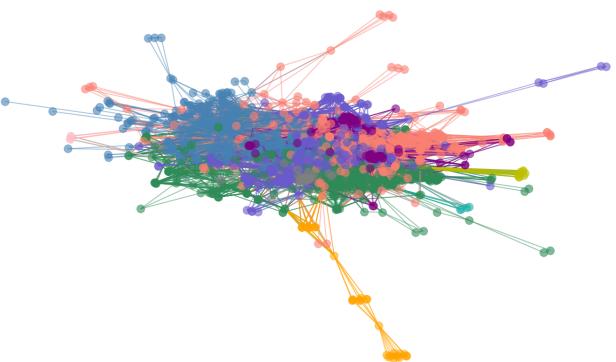


Figura 13: Las comunidades de la red identificadas por color.

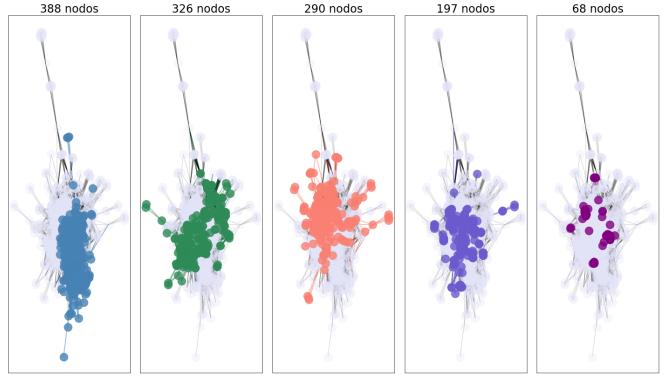


Figura 14: Comunidades 1, 2, 3, 4, 5 de la red.

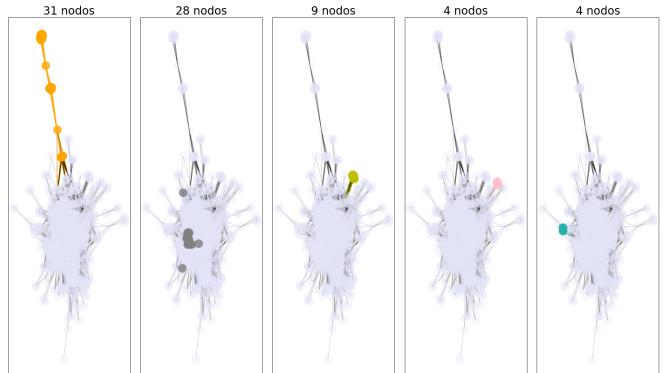


Figura 15: Comunidades 6, 7, 8, 9, 10 de la red.

obtenido un mayor valor para esta medida, y se tiene que el síntoma de *Bleeding gums*, *Ear ache*, *Enlarged or swollen glands* and *Pain when moving eyes* es el más importante.

IV-C. Closeness

Como bien se explica en [10], “la centralidad de *closeness* indica cómo de cerca está un nodo respecto al resto de nodos de la red. Se calcula como la media de los caminos más cortos desde el nodo hasta los demás nodos de la red.“

En la Figura 19 se muestra el histograma de los coeficientes de *closeness* de todos los nodos de la red. Los valores obtenidos no son excesivamente grandes; sin embargo, se puede detectar que es posible alcanzar cualquier nodo de la red de forma sencilla ya que parece que hay muchos caminos

Posición	Síntoma	Eigenvector
1	3	0.13346
2	15	0.13204
3	2	0.12982
4	8	0.12936
5	1	0.12744

Cuadro IV: Los 5 síntomas más importantes según eigenvector.

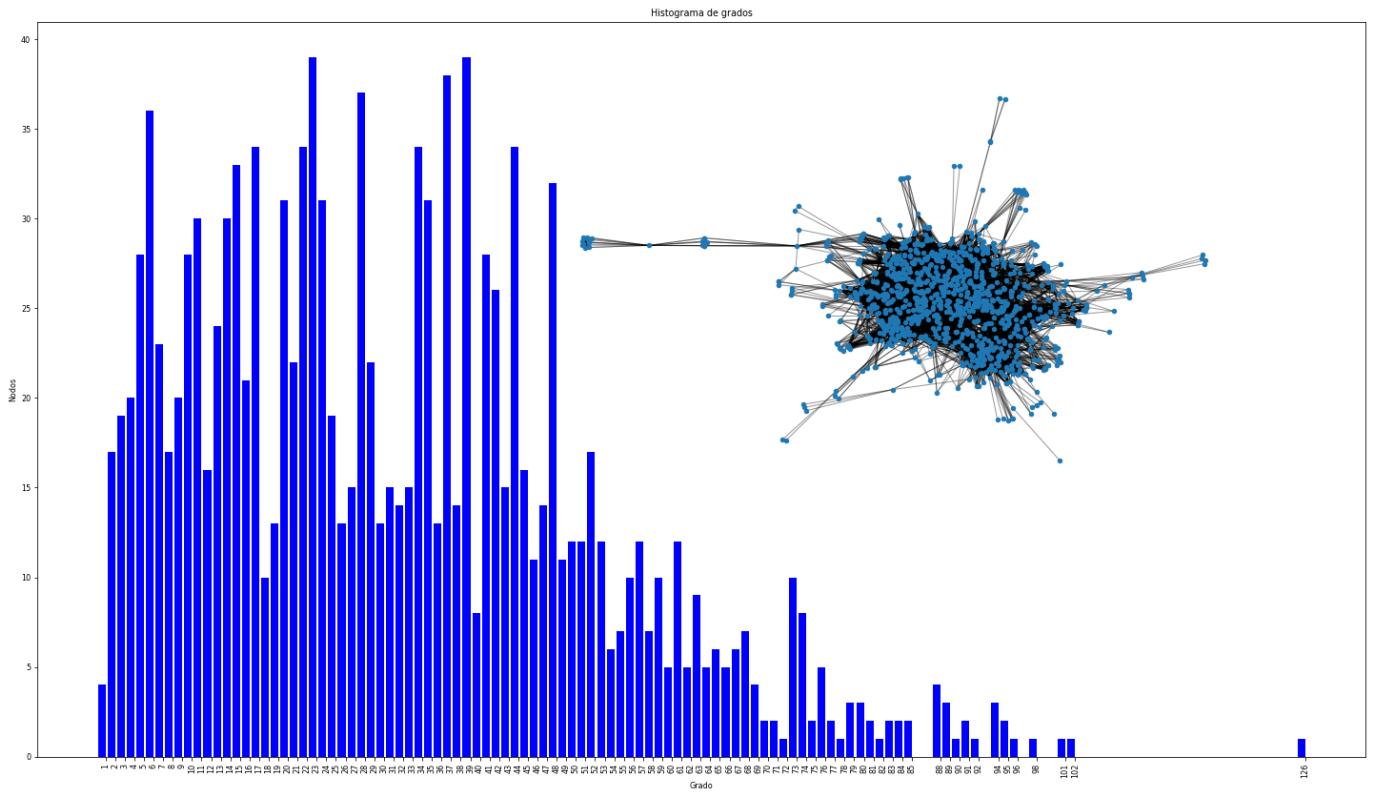


Figura 16: Histograma con los grados de los nodos de la red.

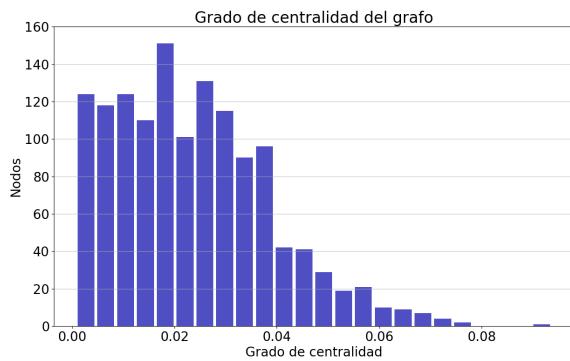


Figura 17: Histograma con los grados de centralidad de los nodos.

cortos dentro de la red.

En el Cuadro V se muestran los 5 síntomas que han obtenido un mayor valor para esta medida, y se tiene que el síntoma de ***Blurry Vision*** es el más importante.

IV-D. Betweenness

Según [11], “la centralidad de *betweenness* analiza cuánto está un nodo en concreto en medio de otros nodos de la red. Esta métrica se mide según el total de caminos cortos que pasan por el nodo, y se considera que un nodo tendrá un alto

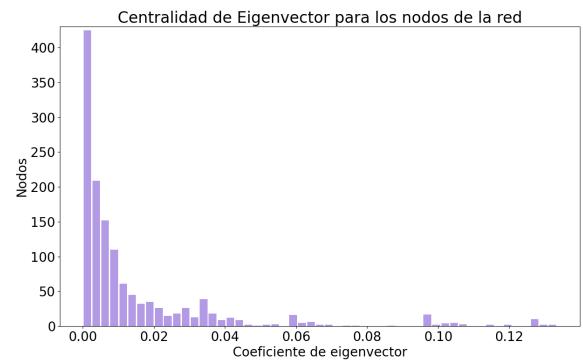


Figura 18: Histograma con los coeficientes de *eigenvector* de los nodos de la red.

Posición	Síntoma	Closeness
1	7	0.42720
2	13	0.42585
3	9	0.42397
4	10	0.41584
5	11	0.41468

Cuadro V: Los 5 síntomas más importantes según closeness.

valor de *betweenness* si aparece en muchos caminos cortos.”

En la Figura 20 se muestra el histograma de los coeficientes de *betweenness* de todos los nodos de la red. La mayoría

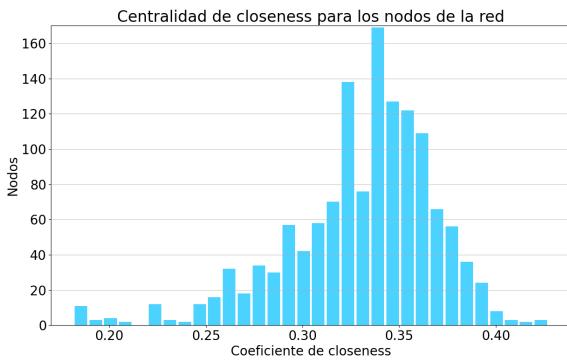


Figura 19: Histograma con los coeficientes de *closeness* de los nodos de la red.

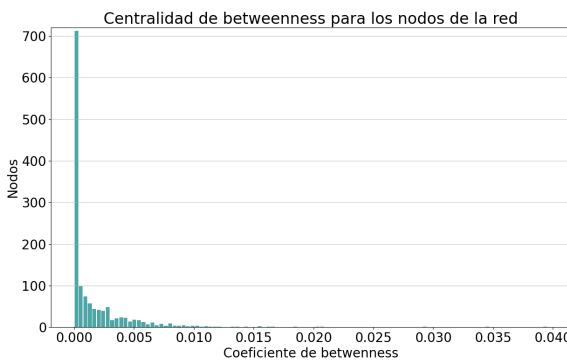


Figura 20: Histograma con los coeficientes de *betweenness* de los nodos de la red.

de nodos ha obtenido valores muy bajos para esta medida. En concreto, 583 nodos tienen un coeficiente de *betweenness* igual a 0. Visualizando la red, se puede apreciar que existe una concentración de nodos y otros nodos que están conectados a esta concentración. Estos nodos se denominan nodos periféricos, y son los que tienen un *betweenness* de 0. En la Figura 21 se muestran dichos nodos.

En el Cuadro VI se muestran los 5 síntomas que han obtenido un mayor valor para esta medida, y se tiene que el síntoma de ***Blurry Vision*** es el más importante.

Posición	Síntoma	Betweenness
1	7	0.03955
2	16	0.03451
3	4	0.02934
4	6	0.02086
5	5	0.02048

Cuadro VI: Los 5 síntomas más importantes según el *betweenness*.

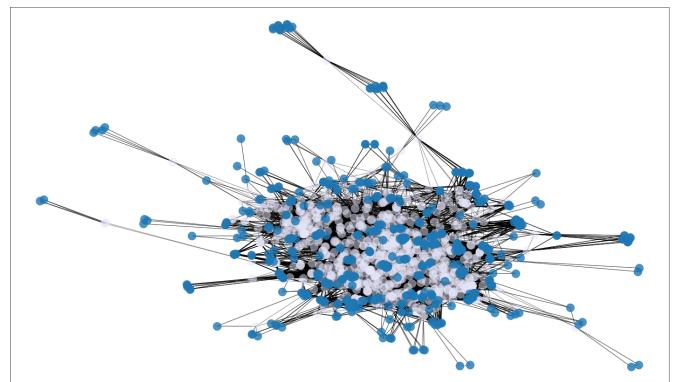


Figura 21: Nodos periféricos con un coeficiente de *betweenness* de 0.

Posición	Centralidad de grados	Eigenvector	Closeness	Betweenness
1	7	3	7	7
2	10	15	13	16
3	12	2	9	4
4	14	8	10	6
5	15	1	11	5

Cuadro VII: Comparación de los resultados obtenidos con los índices de centralidad.

IV-E. Comparación de los resultados

Analicemos ahora los resultados obtenidos para cada índice para determinar si existe alguna correlación entre los resultados obtenidos.

En el Cuadro VII se muestran los 5 síntomas más importantes obtenidos para cada medida empleada. Tras analizar los resultados, podemos determinar que el síntoma de ***Blurry Eyes*** es el más importante según la medida de centralidad de grados, *betweenness* y *closeness*.

Otro síntoma que aparece en el top 5 de dos medidas es el ***Bulging neck veins, Enlarged glands and Pain***.

El resto de nodos que se han detectado como importantes son distintos para cada medida, por lo que no hay mucha correlación entre ellos.

REFERENCIAS

- [1] Symptoms signs a-z list - a. https://www.medicinenet.com/symptoms_and_signs/alpha_a.htm. Accessed: 2020-11-21.
- [2] Qué es el web scraping. <https://aukera.es/blog/web-scraping/>. Accessed: 2020-11-26.
- [3] Beautiful soup documentation. <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>. Accessed: 2020-11-26.
- [4] xc_practical/scraping.py. https://github.com/magheata/xc_practical/blob/main/scraping.py. Accessed: 2020-11-26.
- [5] Fever: Symptoms signs. <https://www.medicinenet.com/fever/symptoms.htm>. Accessed: 2020-11-26.
- [6] Transitivity in a graph. https://transportgeography.org/?page_id=6171. Accessed: 2020-11-26.
- [7] Cores. https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.core.core_number.html#networkx.algorithms.core.core_number. Accessed: 2020-11-30.
- [8] Mason A Porter, Jukka-Pekka Onnela, and Peter J Mucha. Communities in Networks. Technical report.

- [9] Derek L. Hansen, Ben Shneiderman, Marc A. Smith, and Itai Himelboim. Chapter 3 - social network analysis: Measuring, mapping, and modeling collections of connections. In Derek L. Hansen, Ben Shneiderman, Marc A. Smith, and Itai Himelboim, editors, *Analyzing Social Media Networks with NodeXL (Second Edition)*, pages 31 – 51. Morgan Kaufmann, second edition edition, 2020.
- [10] Jennifer Golbeck. Chapter 3 - network structure and measures. In Jennifer Golbeck, editor, *Analyzing the Social Web*, pages 25 – 44. Morgan Kaufmann, Boston, 2013.
- [11] Charles Perez and Rony Germon. Chapter 7 - graph creation and analysis for linking actors: Application to social data. In Robert Layton and Paul A. Watters, editors, *Automating Open Source Intelligence*, pages 103 – 129. Syngress, Boston, 2016.