UNIVERSIDAD NACIONAL DE COLOMBIA

Análisis Estadístico de Redes Sociales: Taller #3

Valentina Cardona Saldaña

Ejercicio #1 y #2

Al revisar los capítulos 6 y 9 de Luke (2015) [1], se comprobó que todos los ejemplos podían replicarse satisfactoriamente. No obstante, es importante destacar que debido a la antigüedad del libro, se encontraron algunas funciones obsoletas, las cuales han sido reemplazadas por nuevas. En tal sentido, se procedió únicamente con la actualización correspondiente. En el Anexo 1 y Anexo 2 se adjuntan las salidas respectivas de este ejercicio. Dado que estos capítulos abordan exhaustivamente cada una de las funciones, no se considera necesario profundizar aquí, prefiriendo focalizar la atención en la siguiente sección 3, que explora una nueva red con mayor detalle.

Ejercicio #3

La base de datos considerada difiere de la del Taller 2, ya que el profesor Juan Camilo Sosa amablemente me proporcionó una base de datos similar, pero enmarcada en el contexto colombiano, la cual era de particular interés para mí. Esta base de datos proviene del trabajo realizado por Samuel Hernando Sánchez Gutiérrez en su tesis de grado, titulada *Modelamiento Bayesiano de redes sociales online de influencia y su impacto en la formación de la opinión pública*, en el marco de su investigación en Análisis de redes sociales. Los datos se encuentran disponibles en `https://github.com/Samuel-col/influenceModel-tesis`. Samuel logró capturar una red de influencia en la plataforma de red social *online Twitter*, durante el debate sobre la Reforma Tributaria, discutida en el Congreso de la República, propuesta por el gobierno colombiano en el segundo semestre de 2022. Samuel recolectó una red dirigida de orden 897 y tamaño 910, con interacciones entre usuarios que estuvieran a favor y en contra del tema de la reforma tributaria. En consecuencia, el usuario emisor es aquel que es retuiteado por el usuario receptor, lo cual indica acuerdo con el contenido del tuit. Sin embargo, no se trabajará con toda la red, sino con la componente gigante, que consta de 634 nodos y 745 enlaces, abarcando el 70% de la red original.

a) Caracterizar la centralidad de los nodos

Para caracterizar la centralidad de los nodos, se comenzó con los grados de salida, ya que estos representan a los usuarios con mayor influencia en la red. El presidente Gustavo Petro obtuvo el mayor número de retuits, seguido por Gustavo Bolívar (ex-Senador de la República y Director del DPS), Wilson Arias (Senador de la República), e Iván Cepeda Castro (Senador de la República); los cuatro son miembros del partido político *Pacto Histórico*, liderado por el actual Presidente. En quinto lugar se encuentra Miguel Uribe Turbay (Senador de la República), un fuerte opositor del gobierno de Gustavo Petro.

Table 1: Grados de entrada

N°	Retuits	Usuario
1	108	petrogustavo
2	45	GustavoBolivar
3	44	wilsonariasc
4	34	IvanCepedaCast
5	31	MiguelUribeT

Adicionalmente, también se calculan medidas de centralidad normalizadas como ejercicio práctico. Estas medidas de centralidad están diseñadas para cuantificar la importancia de los nodos de una red [2]. En primer lugar, la **Centralidad de cercanía** indica que los nodos más relevantes están más cerca (en términos de distancia geodésica) de todos los demás nodos en la red. En términos prácticos, el nodo con mayor centralidad de cercanía es capaz de difundir información rápidamente a través de la red, ya que está cerca de muchos otros nodos. En este caso, el usuario *luiscrh* corresponde al Director de la DIAN, con un $c_{\mathsf{C}}(v) = 0.22$. Por otro lado, la **Centralidad de intermediación** se relaciona con la importancia de un nodo como intermediario en las comunicaciones dentro de la red. El nodo con mayor centralidad de intermediación es *Mamertos0* ($c_{\mathsf{B}}(v) = 9.9986e-06$) que parece ser una cuenta de Twitter de carácter personal. Por último, la **Centralidad propia** se basa en los vectores propios de la matriz de adyacencia de la red. El nodo con mayor centralidad propia es aquel que está conectado con otros nodos que también son importantes; en este caso, el presidente ($c_{\mathsf{E}}(v) = 1$).

Table 2: Centralidad de cercanía, intermediación y propia

N°	Cercanía	Intermediación	Propia
1	luiscrh	Mamertos0	petrogustavo
2	Carlos_toto_n	DavidRacero	wilsonariasc
3	ghitis	laurisarabia	IvanCepedaCast
4	Fabijarabar	MiguelPoloP	GustavoBolivar
5	GrisalesRubio	ELTIEMPO	Fabijarabar

b) Visualizar la centralidad de la red

En este análisis, la **centralidad de intermediación** parece no ser muy informativa para esta red específica; dado que arroja valores muy cercanos a 0. Esto podría deberse a que la red es dirigida o porque la mayoría de los nodos no se encuentran en el camino más corto entre otros pares. La **centralidad de cercanía** y la **centralidad de eigenvector** proporcionan información más valiosa.
La visualización a continuación representa el tamaño de los nodos de acuerdo a sus grados de salida, centralidad de cercanía y centralidad propia. Con esto, podemos evidenciar que las gráficas más reveladoras son las de grado de salida y las de centralidad propia, ya que, a pesar de que los nodos con alta centralidad de cercanía como *luiscrh* podrían ser buenos puntos de difusión de información debido a su proximidad a otros nodos, parece que muchos nodos tienen cercanías muy similares. Por otro lado, los nodos con alta centralidad de eigenvector como *petrogustavo* podrían ser considerados más influyentes debido a sus conexiones con otros nodos importantes (parecidos en tamaño).

Figure 1: Grado de salida

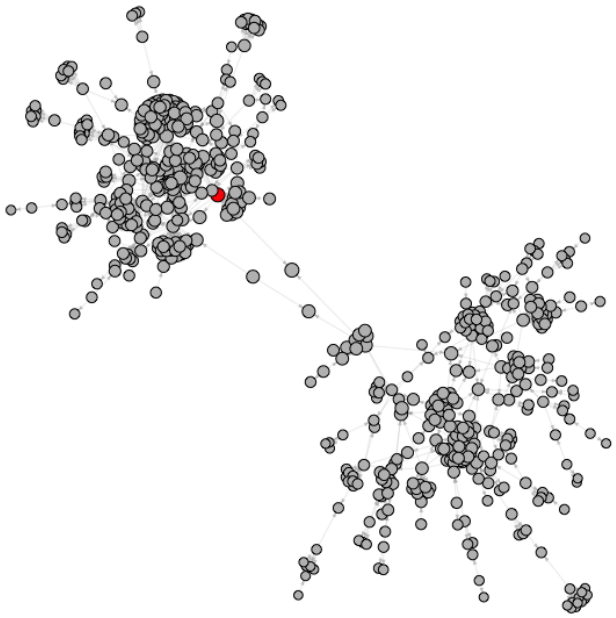


Figure 2: Cercanía

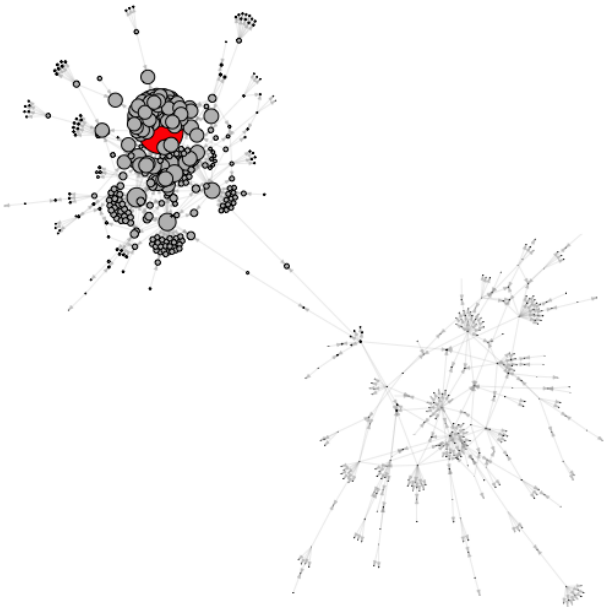


Figure 3: Propia

c) Identificar los puntos de articulación, los puntos aislados y las componentes

Al calcular la *k-connectividad* de la componente gigante, se encontró que es de 0. Esto significa que si se remueve al menos un vértice de la red, la conectividad del subgrafo se vería afectada, volviéndose desconectado. Los **puntos de articulación** de la componente principal suman un total de 111 nodos, lo que corresponde al 17.5% de la red. En la Figura 4 se puede observar que estos puntos de articulación suelen ser los nodos con mayor grado de salida, es decir, los más influyentes en la red.

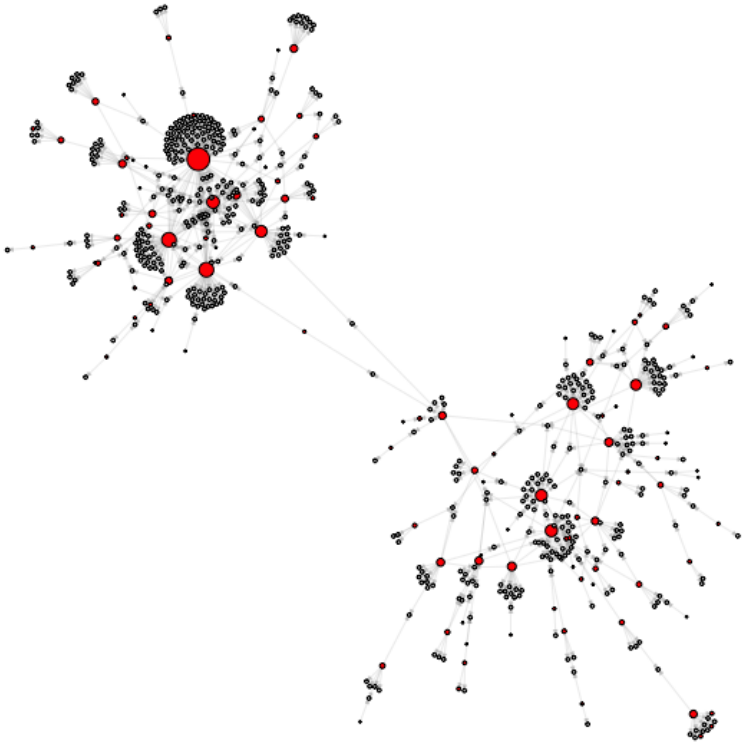


Figure 4: Puntos de articulación

En cuanto a los puntos aislados, esta red no presenta ninguno, teniendo en cuenta su forma de extracción (por interacciones entre usuarios). Asimismo, al extraer la componente gigante, se garantiza el mayor número de nodos conectados, sin puntos aislados. Si consideramos la red original, esta estaba compuesta por un total de 99 componentes, distribuidas como se muestra en la Tabla 4, que indica la cantidad de componentes según el número de nodos que contienen. Se puede concluir que una gran cantidad de los usuarios presentaron díadas y tríadas de interacciones alrededor del debate.

Table 3: Número de nodos por componentes

Número de nodos	**Número de componentes**
2	69
3	17
4	7
5	3
7	1
24	1
634	1

d) Hacer la distribución de las distancia geodésica

Como se ha mencionado anteriormente, la red corresponde a una red dirigida. Su diámetro sería de 2; es decir, el valor de la distancia más grande entre dos nodos en el grafo es 2. Su distancia geodésica promedio es de 1.00534. Respecto a su distribución de distancias, la mayoría de los pares de nodos (745) están a una distancia de 1 paso. Sin embargo, 400573 pares de nodos están no conectados, es decir, el primer vértice no es accesible desde el segundo.

Por otro lado, resulta interesante analizar el caso en que la red fuera no dirigida. El diámetro de la red no dirigida sería de 17. La distancia geodésica promedio en la red no dirigida es de 6.96. En promedio, se

necesitan aproximadamente 7 pasos para llegar de un nodo a otro. La distribución de las distancias muestra que la mayoría de los pares de nodos (36438) están a una distancia de 4 pasos, y no presenta pares no conectados.

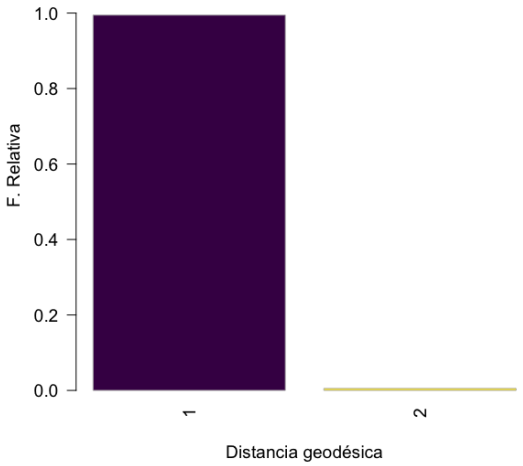


Figure 5: Distribución: Red dirigida

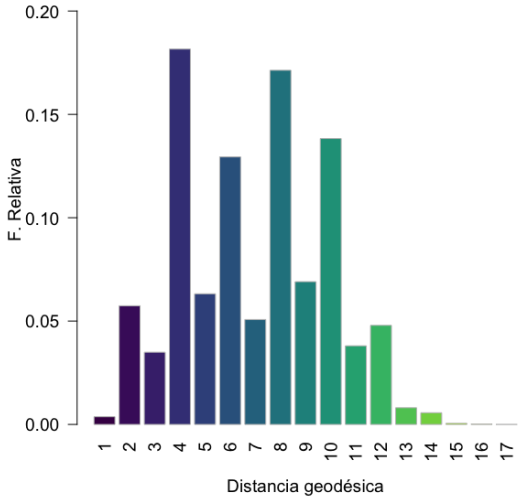


Figure 6: Distribución: Red no dirigida

En general, la red no dirigida tiene un diámetro mayor y una distancia geodésica promedio mayor que la red dirigida; así como su distribución es mucho más dispersa, con una mayor cantidad de pares de nodos a distancias mayores. Esto era en cierta forma esperable; sin embargo, las diferencias son considerables. Puede deberse a la constitución misma de las redes sociales *online*, donde los usuarios que tienen más alcance (mayor número de seguidores) van a conseguir un mayor número de personas que interactúen con ellos y, por ende, una mayor influencia.

e) Determinar si la red es libre de escala

En una red libre de escala, algunos nodos están altamente conectados, es decir, poseen un gran número de enlaces a otros nodos, aunque el grado de conexión de casi todos los nodos es bastante bajo [2]. En la Figura 7 se puede observar que la distribución del grado de entrada parece seguir una ley de potencias, con la mayoría de los nodos teniendo pocos enlaces de entrada y un pequeño número de nodos con muchos enlaces de entrada. La distribución del grado de salida también parece seguir una ley de potencias, aunque con una forma ligeramente diferente a la del grado de entrada. No se puede observar claramente en la figura, pero una gran cantidad de nodos no tienen ningún enlace de salida, es decir, presentan un grado de salida de 0. En conclusión, tanto la distribución del grado de entrada como la del grado de salida son compatibles con la idea de que la red podría ser libre de escala.

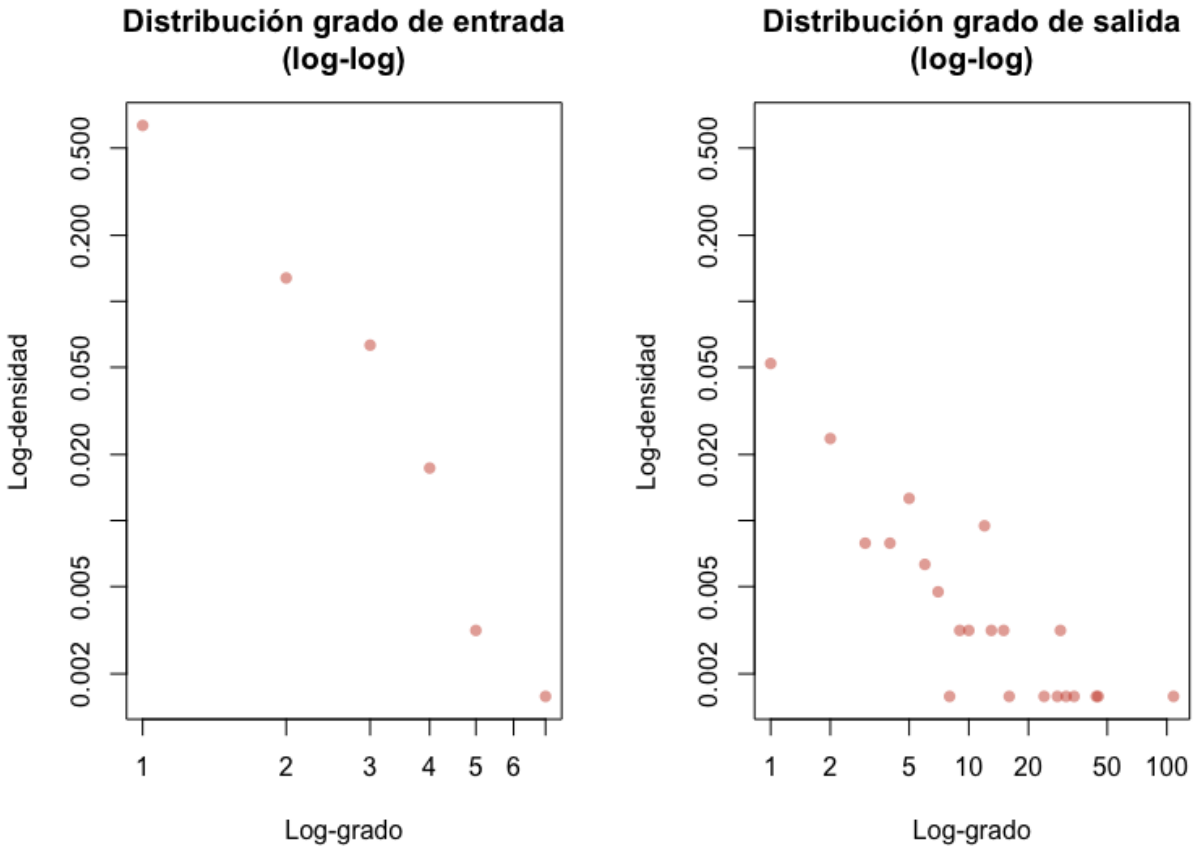


Figure 7: Distribuciones

f) Hacer un censo de los clanes y calcular el número clan

Cabe hacer la claridad de que las funciones utilizadas en esta sección ignoran la dirección de los enlaces al calcular los clanes. Esto significa que la red se trata como no dirigida. Ahora, la siguiente tabla muestra la distribución de tamaños de clanes encontrados en la red. Es de interés considerar solo los clanes de tamaño mayor a 1. Con esto, se observa que 745 clanes tienen un tamaño de 2, y solo hay 1 clan de tamaño 3. Este enfoque resulta interesante para definir la cohesión de una red; a medida que aumenta el tamaño de los clanes, el número de clanes potenciales disminuye drásticamente, lo que sugiere una red con poca estructura comunitaria definida. Por último, el clan más grande identificado (*clan máximo*) tiene un tamaño de 3 e incluye a los usuarios: *Mamertos0*, *RoyBarreras* y *FedericoYBabaji*.

Table 4: Tamaños de clanes

Tamaño	**Número de clanes**
1	634
2	745
3	1

g) Calcular la densidad junto con el coeficiente de agrupamiento de la red

En este caso, la **densidad** del dígrafo es de 0.00186, lo que indica una red escasamente conectada. La mayoría de los nodos no están conectados entre sí y hay un bajo número de enlaces en comparación con el número máximo posible. Por otro lado, el **coeficiente de agrupamiento** (transitividad) mide la tendencia de los

nodos a formar triángulos cerrados. En este caso, el coeficiente de agrupamiento es de 0.000253, lo que indica una baja transitividad en la red. Los nodos que comparten un vecino no tienen una alta probabilidad de estar conectados entre sí. Esto puede deberse, como se mencionó anteriormente, a que la red representa un sistema en el que las relaciones entre los nodos son esporádicas y/o no recíprocas.

h) Particionar la red usando tres métodos de agrupamiento de su elección

Los métodos de agrupamiento utilizados fueron: *Edge-betweenness*, *Walktrap* y *Label propagation*. Estos métodos fueron seleccionados específicamente debido a que operan eficazmente con redes dirigidas, preservando su naturaleza sin convertirlas en redes no dirigidas.

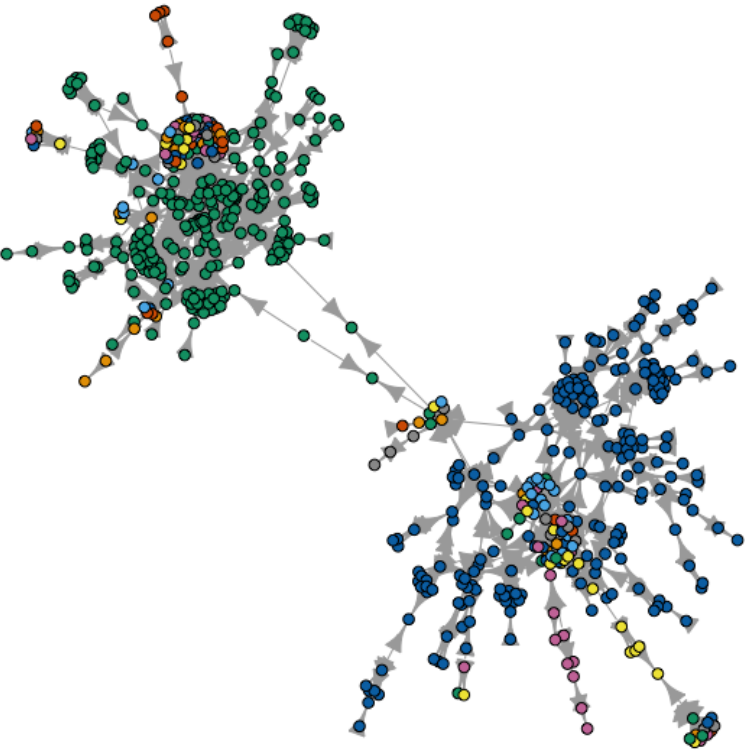


Figure 8: Edge-betweenness

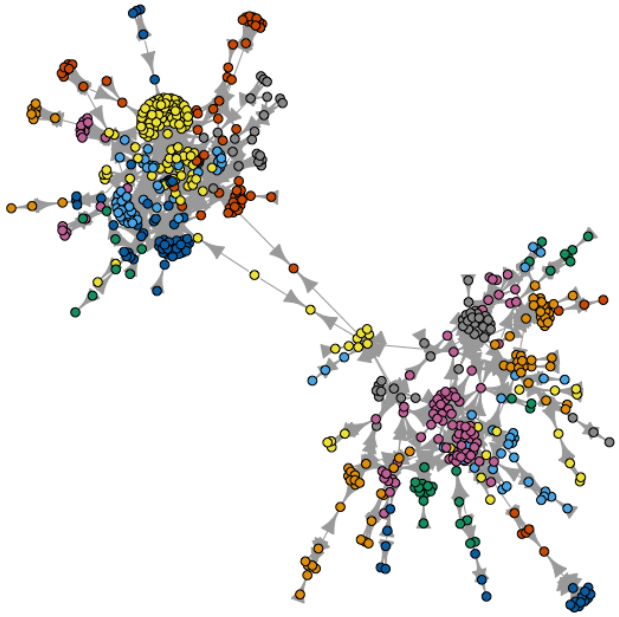


Figure 9: Walktrap

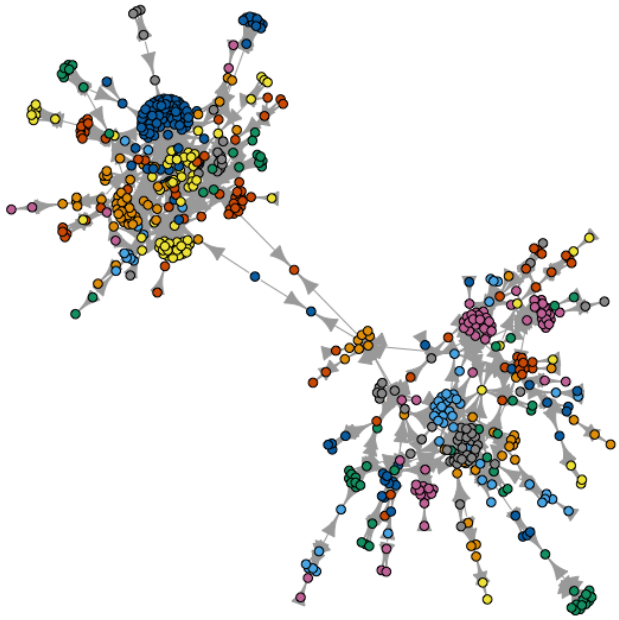


Figure 10: Label propagation

Los métodos de agrupamiento empleados tienen enfoques distintos para identificar comunidades en la red. El método *Edge-betweenness* busca optimizar una métrica de aristas basada en caminos más cortos. Por otro lado, *Walktrap* se fundamenta en la idea de que las caminatas aleatorias cortas tienden a permanecer en la misma comunidad. Mientras tanto, *Label propagation* comienza asignando etiquetas únicas a cada nodo, y los nodos con vecinos que comparten la misma etiqueta tienden a "contagiarse" y adoptar la misma etiqueta, creando así comunidades.
Dadas las características previamente analizadas de la red, parece que el método *Edge-betweenness* es el más adecuado para este contexto. Las aristas con mayor centralidad según este método son aquellas que conectan diferentes comunidades. Además, teniendo en cuenta que la centralidad de los nodos por sus interacciones es de gran relevancia, dado que sus conexiones indican acuerdo en opiniones por parte de los usuarios, este método se alinea bien con los objetivos de análisis de la red.

i) Hacer un análisis de asortatividad de la red

Debido a que la red proporcionada no incluía atributos nodales, se generó uno aleatoriamente para efectos del ejercicio. En este caso, se generó aleatoriamente la variable nodal **Sexo** con las categorías *Femenino* y *Masculino*, donde 308 actores quedaron en la primera categoría y 326 en la segunda. En la Figura 11, el sexo Masculino está representado por el color *naranja* y el sexo Femenino por el color *amarillo*.

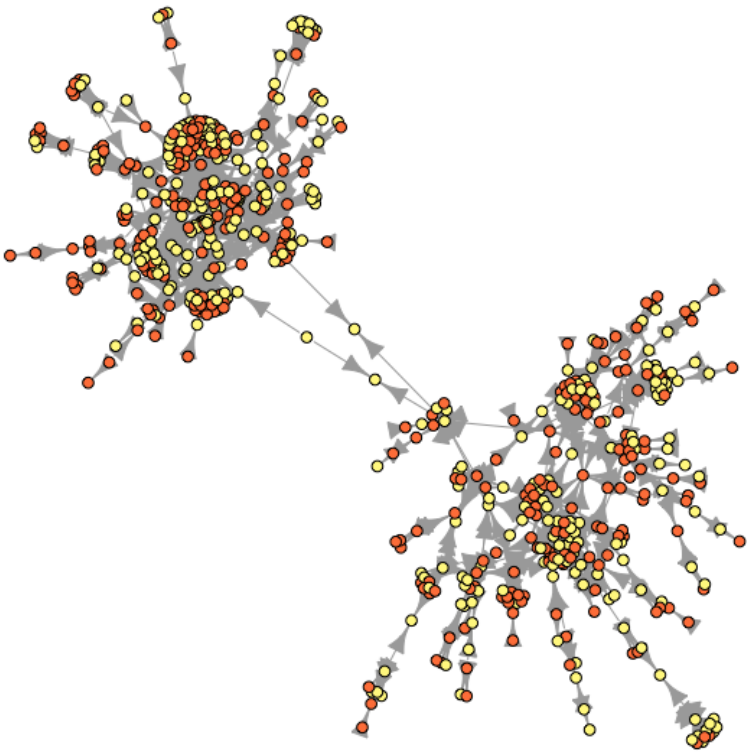


Figure 11: Características nodales

La **asortatividad nominal** mide la tendencia de los nodos a conectarse con otros nodos que comparten el mismo atributo. En este caso, el atributo es el sexo. El valor de 0.044 indica una asortatividad nominal muy cercana a 0. Esto significa que los nodos no tienden a conectarse preferentemente con otros del mismo sexo ni con nodos del sexo opuesto. Esto era esperable dado que los datos se generaron aleatoriamente.

Por otro lado, la **asortatividad de grado** mide la tendencia de los nodos a conectarse con otros nodos que tienen un grado similar. Un valor positivo indica que los nodos con un alto grado tienden a conectarse con otros nodos con un alto grado, mientras que un valor negativo indica una tendencia a conectarse con nodos con un grado bajo. El valor de -0.0984133 indica que no hay una tendencia clara de conexión entre nodos basada en el grado, dado que es muy cercano a 0.

j) Interpretar los resultados

Después de realizar un exhaustivo análisis descriptivo de la red, podemos concluir que presenta una serie de características distintivas que reflejan la dinámica y la estructura de las interacciones entre los usuarios en el contexto del debate sobre la Reforma Tributaria en Colombia. La red exhibe una distribución de grados heterogénea, con algunos nodos altamente conectados y una gran cantidad de nodos con conexiones más limitadas. Además, se observa una tendencia hacia la formación de dos comunidades políticas, aunque estas no son muy definidas. La centralidad de los nodos y la identificación de comunidades han revelado

la presencia de usuarios influyentes y la existencia de agrupaciones en posturas respecto al tema de interés dentro de la red.

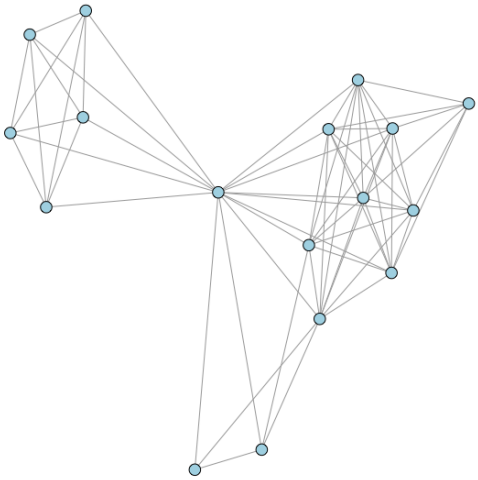
References

[1] Douglas A Luke. *A user's guide to network analysis in R*, volume 72. Springer, 2015.

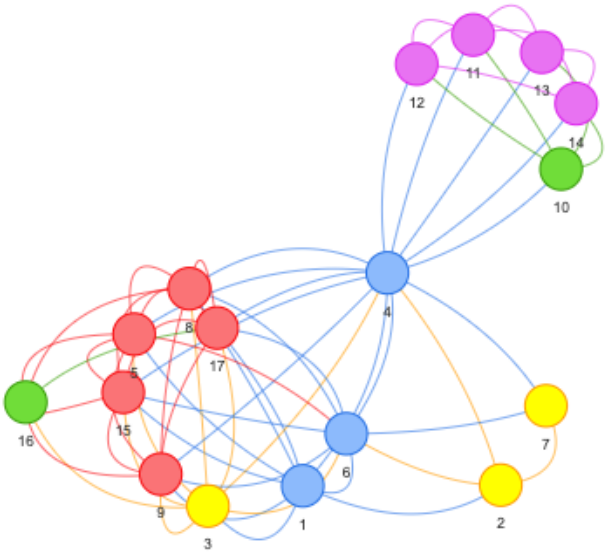
[2] Juan Camilo Sosa. Caracterización de vértices.

Anexo 1: Salidas Capítulo 6

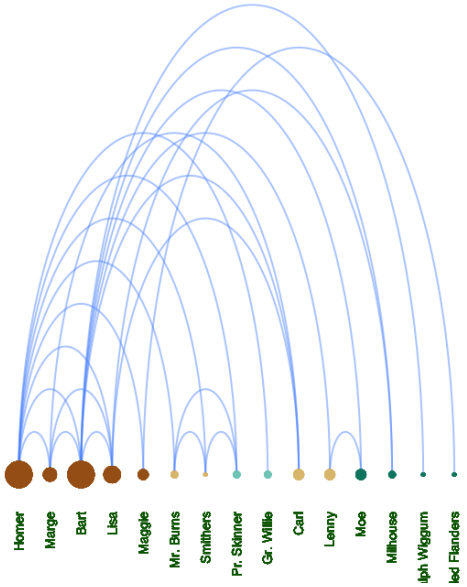
6.1.1 Simple Interactive Networks in igraph



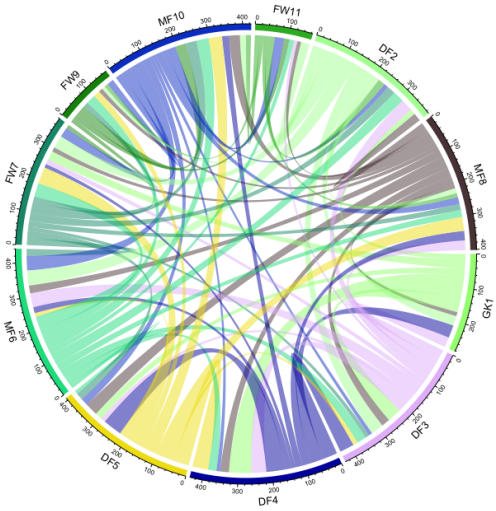
6.1.2 Publishing Web-Based Interactive Network Diagrams

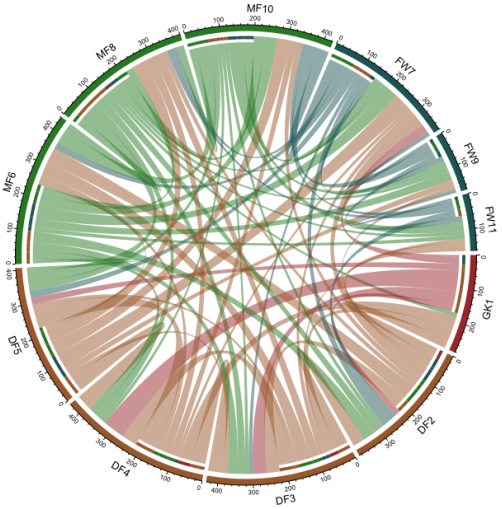


6.2.2 Chord Diagrams

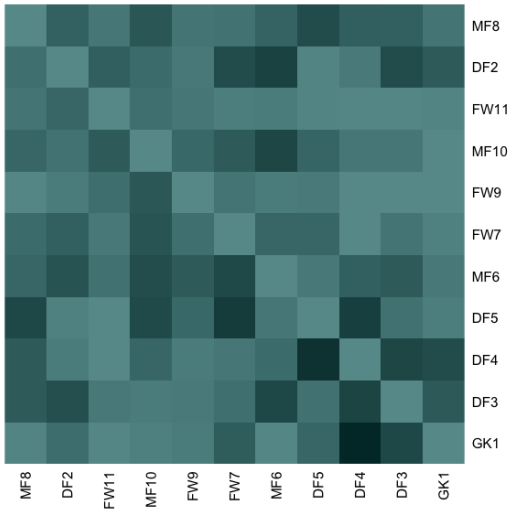


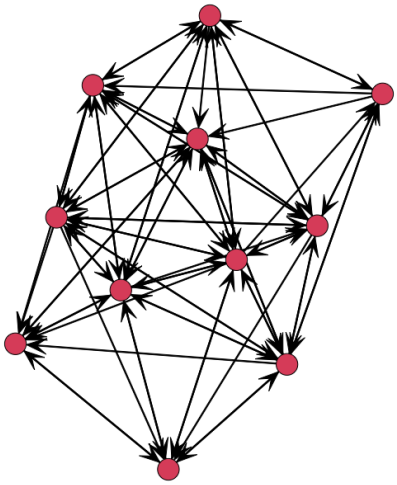
6.2.3 Heatmaps for Network Data

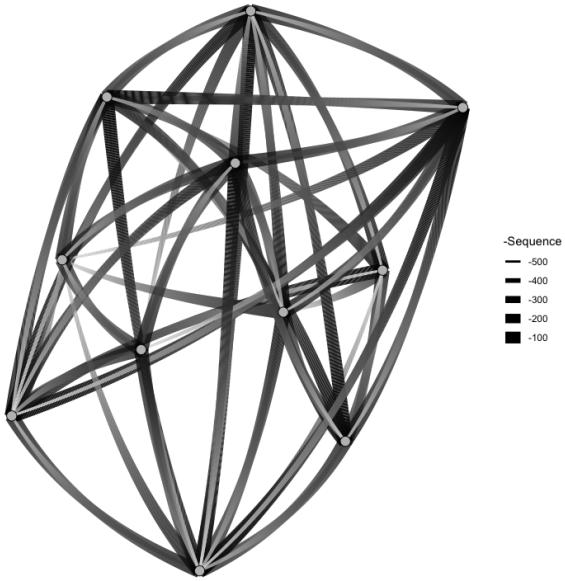




6.3.1 Network Diagrams with ggplot2

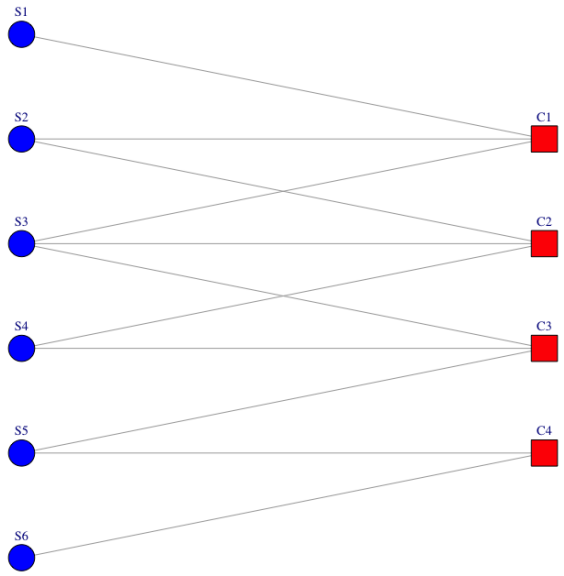




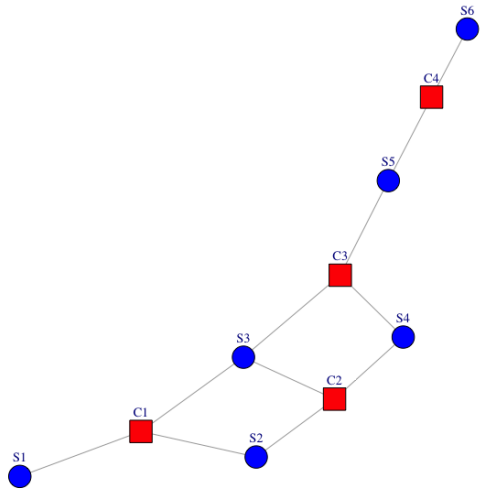


Anexo 2: Salidas Capítulo 9

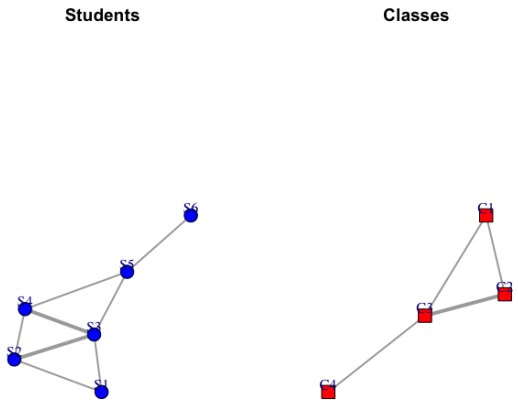
9.1.2 Bipartite Graphs



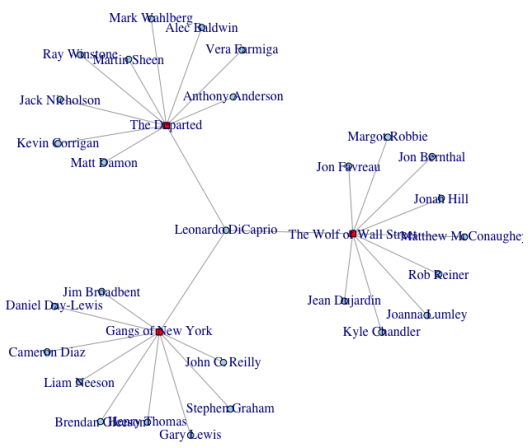
9.2.1 Creating Affiliation Networks from Incidence Matrices

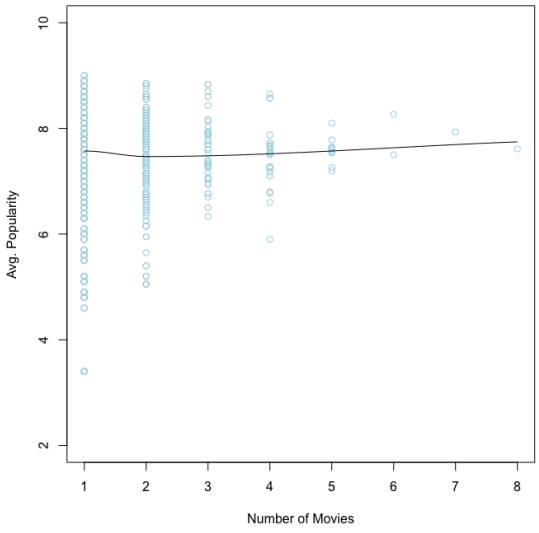


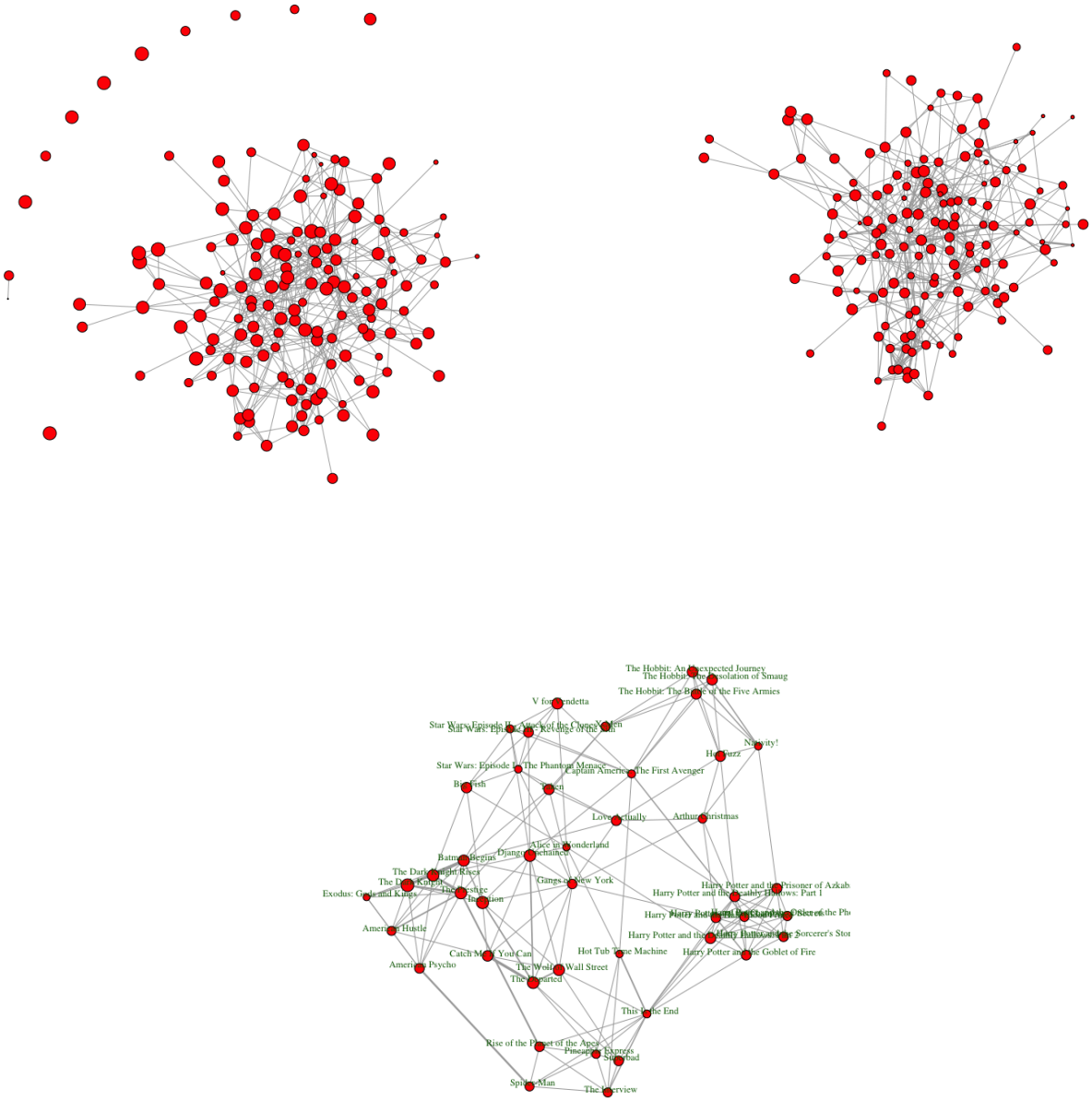
9.2.4 Projections



9.3.1 Analysis of Entire Hollywood Affiliation Network






The Hobbit: An Unexpected Journey
The Hobbit: The Desolation of Smaug
The Hobbit: The Battle of the Five Armies
V for Vendetta
Star Wars: Episode III - Revenge of the Sith
Nativity!
Hot Fuzz
Star Wars: Episode I - The Phantom Menace
Captain America: The First Avenger
Love Actually
Arthur Christmas
Alice in Wonderland
Django Unchained
Batman Begins
The Dark Knight Rises
Gangs of New York
Harry Potter and the Prisoner of Azkaban
Harry Potter and the Deathly Hallows: Part 1
Exodus: Gods and Kings
The Prestige
Inception
American Hustle
Harry Potter and the Sorcerer's Stone
Hot Tub Time Machine
Harry Potter and the Goblet of Fire
Catch Me If You Can
American Psycho
The Wolf of Wall Street
The Departed
This Is the End
Rise of the Planet of the Apes
Pineapple Express
Sinbad
Spider-Man
The Interview