PEC-1

Código de alumno: Polaris

30 de mayo de 2022

Introducción

En este trabajo se ha estudiado la red formada por los concejales de la actual legislatura del ayuntamiento de Madrid. El estudio se ha hecho a partir de los datos recogidos en la red social Twitter. Se ha considerado a cada concejal como un nodo, y los tweets en los que un concejal menciona a otro son las aristas dirigidas. Este trabajo tiene tres objetivos.

El primero es estudiar cómo interactúan los concejales en Twitter y las diferencias (si las hay) en en función del partido al que pertenezcan. Para ello se estudiarán las métricas de grado de entrada y salida, y la centralidad de vector propio agregadas por partido.

El segundo es comprobar si agrupando los nodos por clase de modularidad o usando aprendizaje no supervisado el resultado se parece a la agrupación por partido político.

El último objetivo es estudiar la homofilia, tanto de la red completa como la de las subredes que surgen al agrupar por aprendizaje no supervisado, modularidad y partido.

Descripción de los datos

Para crear la red se han utilizado tweets de los concejales desde el día 15 de Junio de 2019, que fue cuando empezó la actual legislatura. De los 57 concejales en el ayuntamiento de Madrid solo 50 tienen cuenta en Twitter, y de esos 50 hay 3 que nunca han interactuado con los demás y no han sido incluidos en la red. Por lo tanto la red consta de 47 nodos.

Para buscar los datos en Twitter se ha utilizado la herramienta Twint. Con Twint se han obtenido todos los tweets de los concejales que mencionan a otros concejales desde el comienzo de la legislatura. A partir de estos tweets se ha construido una red dirigida, en la que el origen de las aristas es la persona que escribe el tweet, y el objetivo la persona a la que se menciona en el mismo. El peso de la arista es igual al número de veces que un concejal ha mencionado a otro.

El resultado final es una red con 47 nodos y 710 aristas, que se puede ver en la figura 1.

Análisis estático

Para empezar, se ha estudiado el grado de salida ponderado. En la figura 2 se observa un histograma de esta métrica. Se puede observar que la mayoría de los nodos tienen un grado de salida ponderado relativamente pequeño, lo que nos indica que la mayoría de concejales no interactúan mucho con sus colegas en Twitter. Hay un nodo outlier con un grado de salida de 613 que corresponde al concejal Felix López-Rey Gómez.

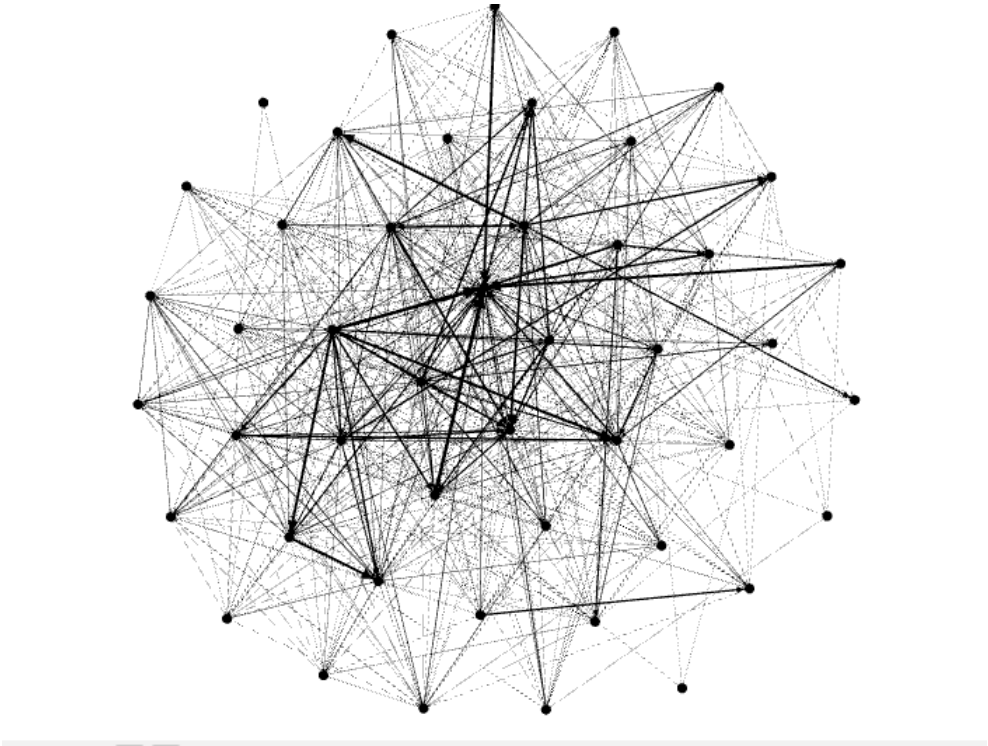

Figura 1: Red de concejales con distribución Fruchterman Reingold

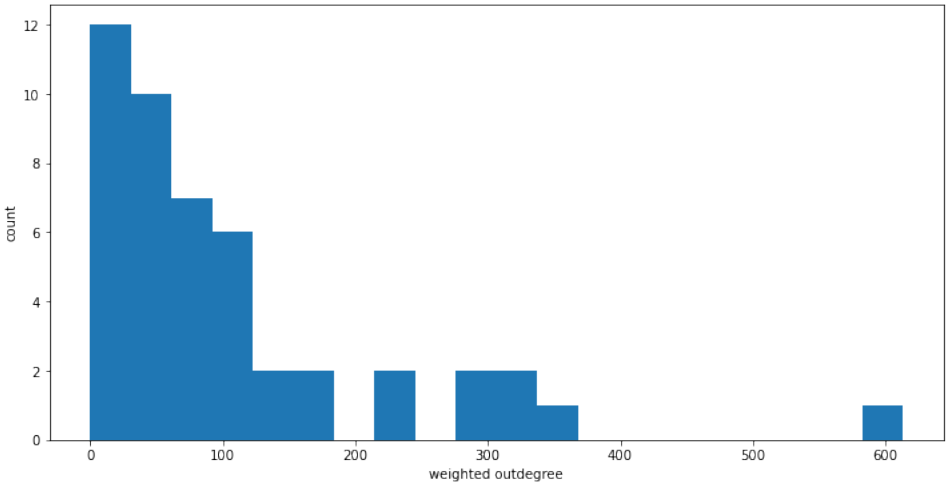

Figura 2: Histograma de el grado de salida ponderado

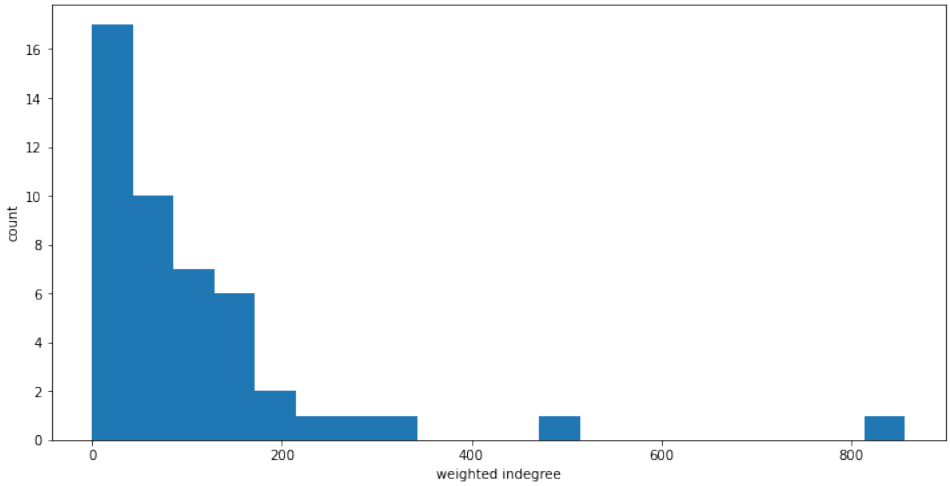

Figura 3: Histograma de el grado de entrada ponderado

En la figura 3 se muestra un histograma del grado de entrada ponderado. De nuevo se observa que la mayoría de nodos tienen un valor relativamente bajo, lo que indica que la mayoría de concejales no suelen ser mencionados por sus colegas. En este caso tenemos dos outliers, uno con grado de entrada de 857 y otro con 487. Estos dos nodos corresponden a José Luis Martínez-Almeida y Begoña Villacís respectivamente. Es bastante lógico que estas dos personas sean las más mencionadas ya que son el alcalde y la vicealcaldesa de Madrid.

En la figura 4 se muestra el histograma de la centralidad del vector propio. Aquí se ve una distribución más homogénea que en los dos histogramas anteriores, con picos en los valores de 0.2 y 0.4. De nuevo hay un outlier con valor de centralidad de 1. Este nodo corresponde a José Luis Martínez-Almeida, lo cual tiene sentido ya que al ser el alcalde es normal que sea un nodo muy importante en la red.

Puesto que Martínez-Almeida es el nodo con mayor centralidad y con mayor grado de entrada, se he considerado que exista una correlación entre estas dos métricas de la red. Para comprobar si existe dicha correlación se calculo

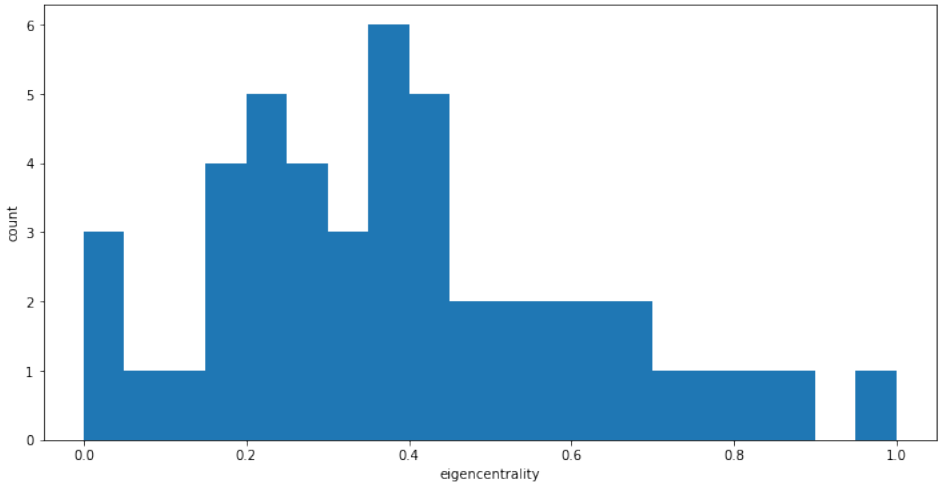

Figura 4: Histograma de la centralidad de vector propio

el coeficiente de Pearson que dio un resultado de 0.77. También se ha representado un scatterplot del grado de entrada frente a la centralidad del vector propio en la figura 5.

Viendo el valor del coeficiente de Pearson y el scatterplot se puede concluir que si que hay una correlación entre ambas cantidades.

Análisis por partido

A continuación se estudian las mismas métricas que en el apartado anterior pero agregadas por partido político. Para empezar, en la figura 6 se muestra la media de el grado de salida ponderado de cada partido. Se puede observar que los partidos con valores más altos son el PSOE y Más Madrid, los principales partidos de la oposición.

En la figura 7 se muestra el grado de entrada ponderado promedio. En este caso, los valores más altos son los de Ciudadanos y el PP, los partidos que gobiernan el ayuntamiento.

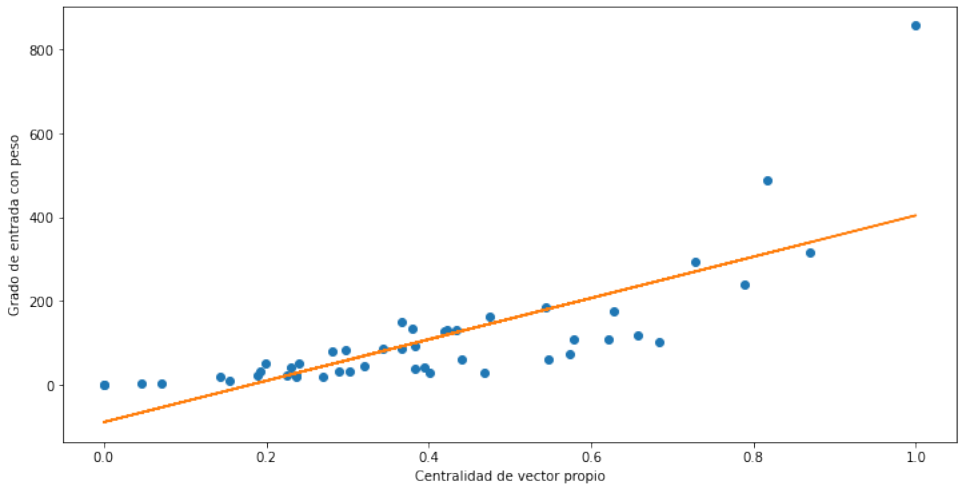

Figura 5: Scatterplot de grado de entrada ponderado frente a la centralidad del vector propio

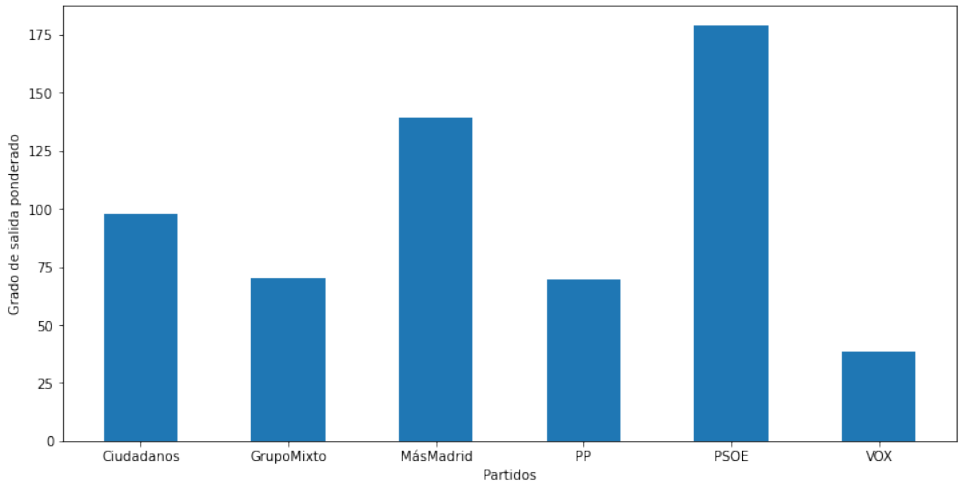

Figura 6: Grado de salida ponderado promedio de cada partido

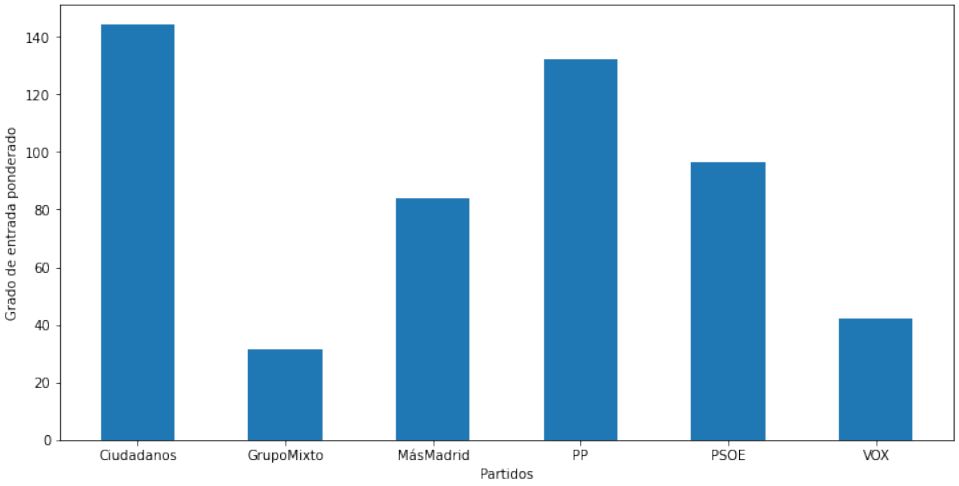

Figura 7: Grado de salida ponderado promedio de cada partido

Por último, la figura 8 muestra la centralidad de vector propio promedio. De nuevo el PP y Ciudadanos tienen los valores más altos. Esta métrica representa la importancia de un nodo en la red, así que es lógico que los nodos de los partidos gobernantes tengan un valor promedio mayor que el resto.

Clasificación por clase de modularidad y por aprendizaje no supervisado

El ayuntamiento de Madrid está dividido en 6 partidos políticos: PP, Ciudadanos, PSOE, Más Madrid, Vox y el grupo mixto. Puesto que en el enunciado de la práctica se pide agrupar los nodos por clase de modularidad y con apren-

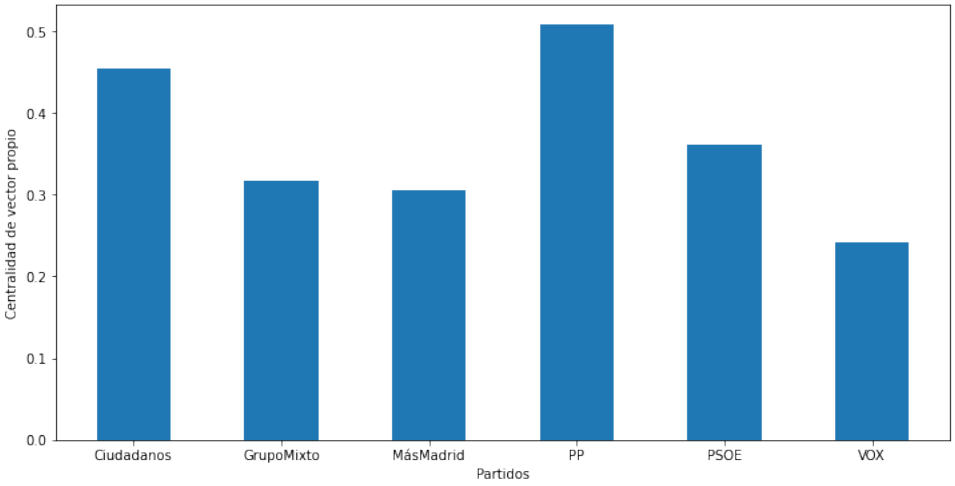

Figura 8: Centralidad del vector propio promedio de cada partido

dizaje automático, se ha decidido que es interesante ver si estas clasificaciones se asemejan a la división por partido político.

Para hacer la clasificación por clase de modularidad se ha utilizado Gephi y se ha seguido el mismo procedimiento explicado en la tarea complementaria 2 del tema 2. Para obligar a Gephi a que salieran 6 grupos distintos se escogió una resolución de 0.7.

Para formar los clústeres por aprendizaje no supervisado se utilizó la librería Sklearn de Python. El modelo escogido fue K-Means por ser un método fácil de implementar y de ejecución rápida. Aunque según los coeficientes de silueta el número óptimo de clústeres es 2, se decidió dividir en 6 para que coincidiera con el número de partidos.

En la figura 9 se muestra un mapa de calor que representa la coincidencia entre partidos políticos y clases de modularidad. Se puede ver que hay partidos que corresponden muy bien con clases de modularidad, como el PP cuyos miembros caen todos en la clase 0, o Más Madrid, que tiene el 92 % de sus miembros en la clase 2. Otros partidos como el PSOE o Vox también

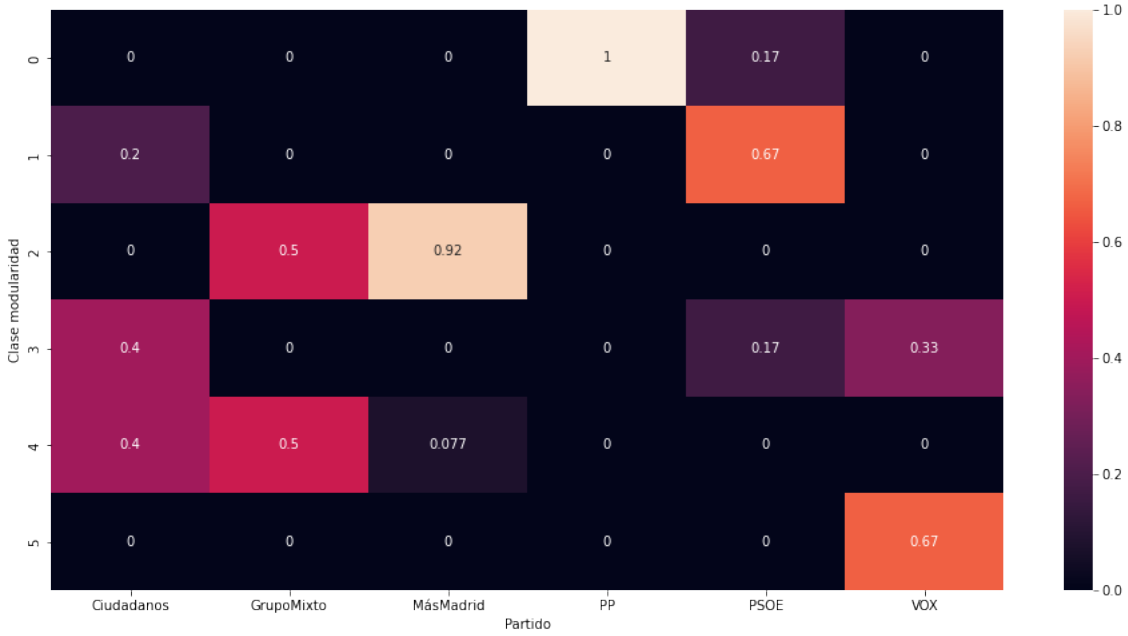

Figura 9: Mapa de calor de la clase de modularidad frente al partido político. Las columnas están normalizadas para que la suma de sus elementos sea 1.

tienen la mayoría de sus nodos en una clase concreta (en la 1 y en la 5 respectivamente). Por último, los nodos de Ciudadanos y el grupo mixto están repartidos entre varias clases sin predominar ninguna.

En la figura 10 se ve un mapa de calor que representa la coincidencia entre clústeres y clases de modularidad. Los clústeres 2 y 5 tienen la totalidad de sus nodos respectivamente en las clases 0 y 2. El resto de clústeres tienen sus integrantes bastante repartidos. Se puede observar que la clase de modularidad 5 está prácticamente vacía.

Por último en la figura 11 se muestra un mapa de calor que representa la coincidencia entre clústeres y partidos políticos. Aquí se ve que la mayoría de partidos están en el clúster 0, mientras que el resto están bastante vacíos.

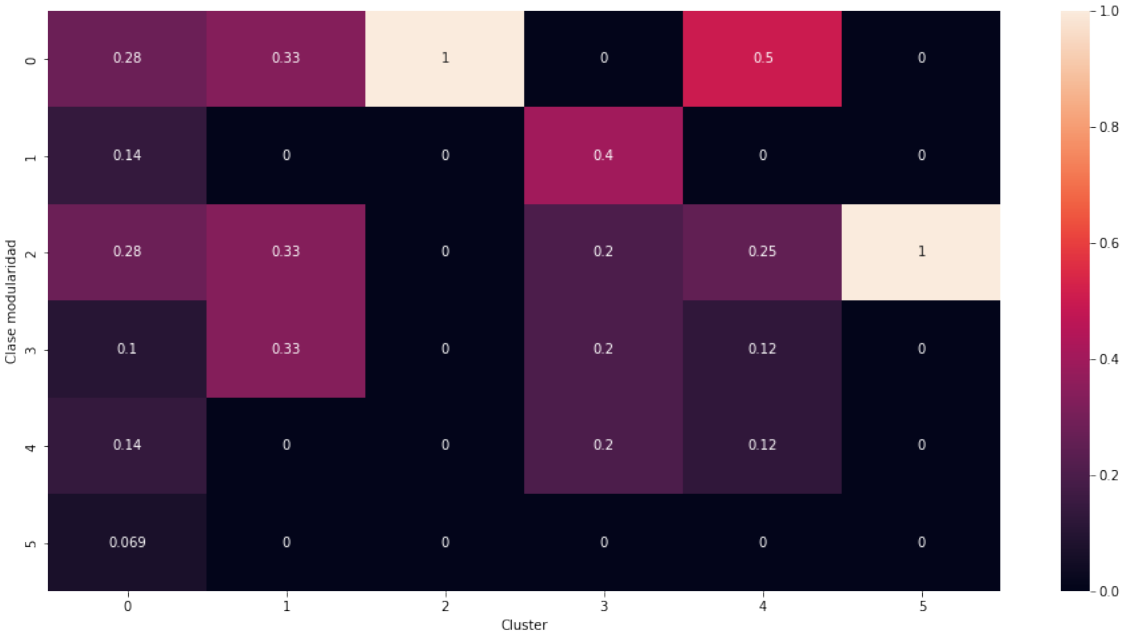

Figura 10: Mapa de calor de la clase de modularidad frente al clúster. Las columnas están normalizadas para que la suma de sus elementos sea 1.

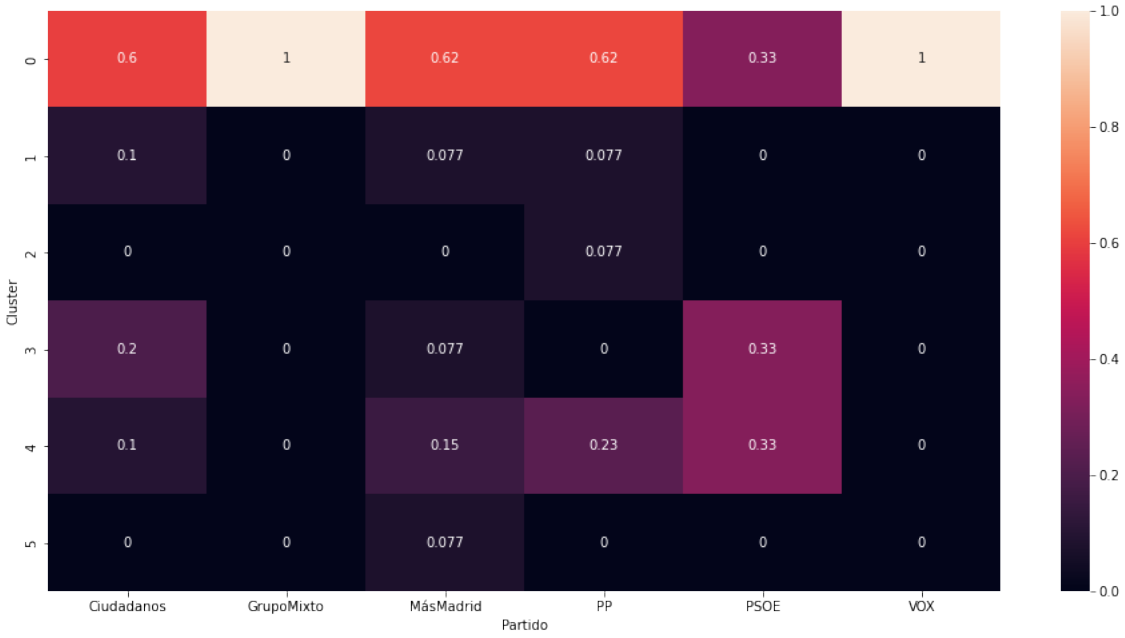

Figura 11: Mapa de calor de clúster frente a partido político. Las columnas están normalizadas para que la suma de sus elementos sea 1.

Estudio de la similaridad

Para estudiar la similaridad se usará la equivalencia estructural, que mide la similaridad en función de el número de vecinos coincidentes entre dos nodos. Al ser una red de 47 nodos, la matriz de adyacencia resultante tiene 2209, y verlo representado en un mapa de calor no es muy informativo. Por lo tanto, se usarán los promedios de los valores de la matriz de adyacencia. Para calcular este promedio se utilizan solo los valores por encima de la diagonal principal de la matriz de adyacencia, ya que la matriz es simétrica, y la diagonal principal son todo 'unos' debido a que los nodos son idénticos a sí mismos. Usando este procedimiento, se calculó le similaridad promedio de la red completa que es 0.27.

Primero se discutirá la división en subredes por partido. Se dividió la red completa en 6 subredes, de manera que cada una solo contuviera nodos de un único partido. De cada subred se calculó la matriz de adyacencia, y de cada matriz se calculó el valor promedio. El resultado puede observarse en la figura 12.

Lo primero que llama la atención en la figura es que hay 3 partidos que tienen similaridad negativa. En el grupo mixto no es extraño, ya que está formado por concejales de distintos partidos que no tienen suficiente representación como para formar grupo propio. Si que es más sorprendente en el PSOE y en Vox. Esta similaridad negativa indica que los concejales de un mismo partido pueden interactuar con grupos de personas totalmente distintas entre si. Más Madrid, PP y Ciudadanos tienen unas similaridades promedio cercanas a la de la red completa.

A continuación se discute la división de la red por clase de modularidad. En la figura 13 se muestra el promedio de similaridad de cada subred.En este caso todos los valores son positivos y están cerca de el valor de la red completa.

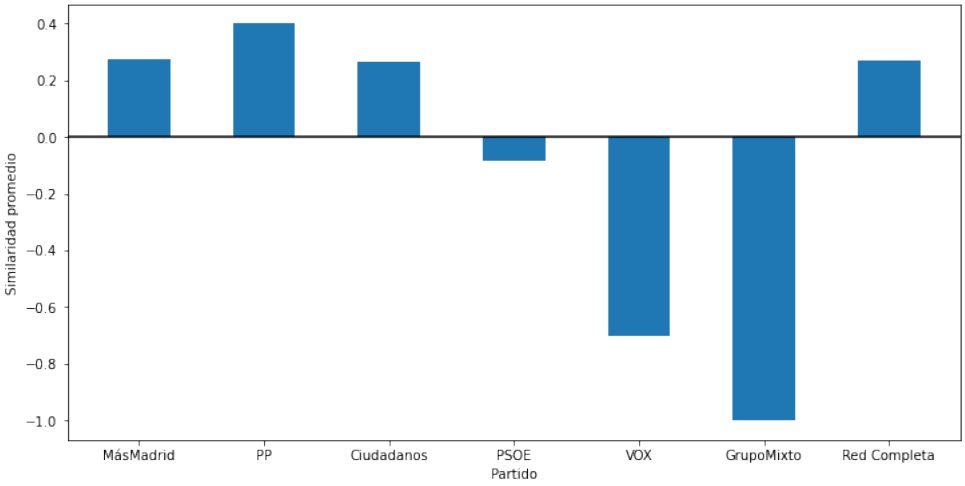

Figura 12: Similaridad promedio de cada subred haciendo la división por partidos. Se incluye el promedio de las similaridades de la red completa para comparar.

Esto es debido a que al hacer la división por modularidad se intenta que todas las subredes tengan una similaridad alta. La clase número 5 no se ha incluido en la figura ya que todos los valores eran 'Not a Number'.

Por último, se discute la división de la red por clúster, cuyo resultado se ve en la figura 14. En este caso las similaridades son negativas o muy cercanas a 0. Esto indica que el algoritmo K-Means no es bueno a la hora de formar clústeres con nodos similares. Los clústeres 2 y 3 no se han incluido en la gráfica porque todos los valores eran 'Not a Number'.

Estudio dinámico de la similaridad

Para estudiar como evolucionaba la red a lo largo del tiempo se ha estudiado cómo varía la similaridad en cada mes. Se ha calculado la media de los valores de la matriz de adyacencia, como en el apartado anterior, pero teniendo en cuenta solo las interacciones que han tenido lugar en ese mes. El resultado

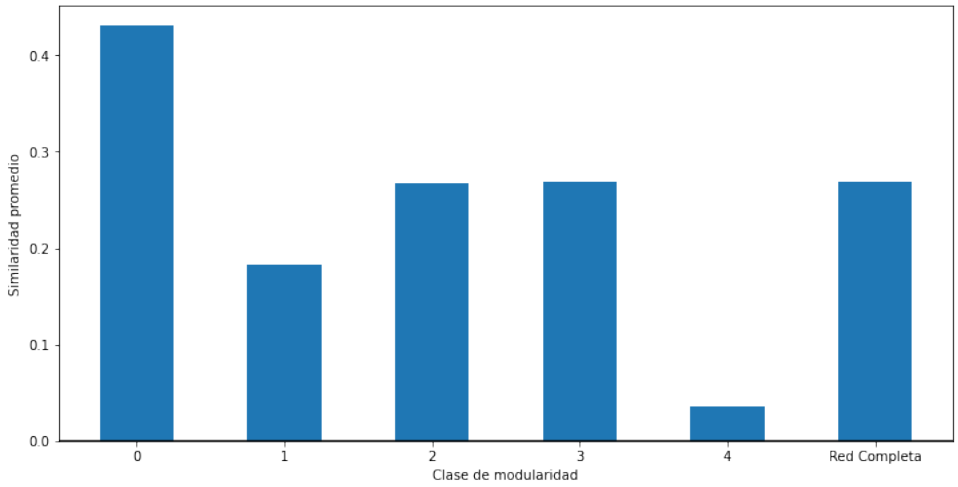

Figura 13: Similaridad promedio de cada subred haciendo la división por clase de modularidad. Se incluye el promedio de las similaridades de la red completa para comparar.

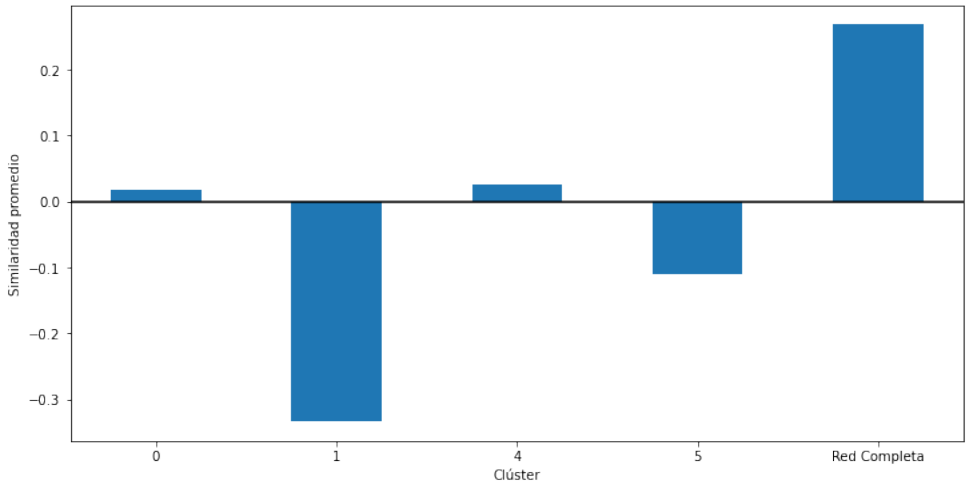

Figura 14: Similaridad promedio de cada subred haciendo la división por clúster. Se incluye el promedio de las similaridades de la red completa para comparar.

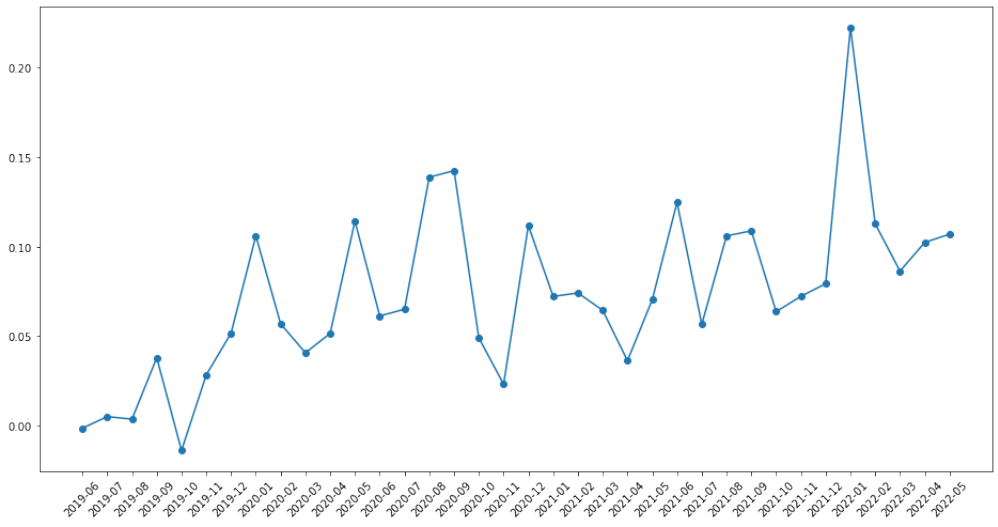

Figura 15: Media de la similaridad de la red en cada mes

se muestra en la figura 15. A primera vista parece que la similaridad cambia muy bruscamente algunos meses, aunque en realidad estos cambios suelen ser pequeños, de intervalos de 0.05 aproximadamente. El valor mínimo (0.00) se encuentra al empezar la legislatura (Junio de 2019) y el máximo (0.22) en Enero de 2022. A pesar de que los cambios son pequeños, si que se aprecia una ligera tendencia a que la similaridad aumente en el tiempo.

Conclusiones

Estudiando el grado de salida y el de entrada de los distintos partidos, se ha observado que los que gobiernan el ayuntamiento (PP y Ciudadanos) son los que tienen mayor grado de entrada, mientras que los principales de la oposición (Más Madrid y PSOE) tienen el mayor grado de salida. Esto indica que gran parte de las interacciones de esta red tienen como origen un nodo de un partido de la oposición y como objetivo un nodo de un partido

del gobierno. Una posible explicación es que muchos de estos tweets sean ataques o quejas de los partidos de la oposición hacia el ayuntamiento.

También se ha comprobado que existe una correlación entre la centralidad del vector propio y el grado de entrada. Al ser la centralidad del vector propio una medida de la importancia del nodo, es lógico que los nodos más importantes sean los que más interacciones reciben. Prueba de esto es que los nodos con mayor grado de entrada y centralidad son el alcalde y la vicealcaldesa.

Al agrupar los nodos por clase de modularidad se ha comprobado que algunos de los grupos formados coinciden bastante bien con algunos partidos políticos. En particular, casi todos los nodos de Más Madrid estaban en la clase de modularidad 2 y todos los del PP en la 0. Por otro lado, la agrupación mediante K-Means ha formado clústeres que no tienen mucha similitud ni con las clases de modularidad ni con los partidos políticos. Es posible que esto se deba a que se decidió agrupar en 6 clústeres cuando los coeficientes de silueta indicaban que el número óptimo era 2.

En el estudio de la similaridad, se han observado grandes diferencias en función de la manera de dividir en subredes. Al dividir por partidos, se ha visto que algunas redes tenían una similaridad promedio alta, como la de el PP y Más Madrid, mientras que otras presentaban valores negativos. El hecho de que el PP y Más Madrid tengan similaridad promedio relativamente alta y que encajen bastante bien en clases de modularidad concretas indica que estos partidos son muy homogéneos.

Al estudiar la similaridad de las distintas clases de modularidad se han obtenido valores cercanos a los de el promedio de la red total. En cambio, los clústeres formados por K-Means tienen similaridades promedio muy bajas o negativas, lo que indica que este modelo no es bueno en crear clústeres con nodos similares entre si.

Por último, en el estudio dinámico de la similaridad promedio se ha observado que esta crece ligeramente con el paso del tiempo.