Laboratorio 2

Estructura de las grandes redes

Grupo

Sara Silva - (5.150.913-2) - sara.silva@fing.edu.uy
Tomás Vázquez - (5.020.594-1) - tomas.vazquez@fing.edu.uy

Instituto de Ingeniería Eléctrica - Facultad de Ingeniería
Universidad de la República
2023

Distribución de grados

Para comenzar se implementó una función que permita calcular la distribución de grados de un grafo, a partir de la lista de los grados de los nodos del mismo. Esta función implementada en el archivo .ipynb, fue probada para el grafo de la Figura 1, y el histograma obtenido fue el de la Figura 2.

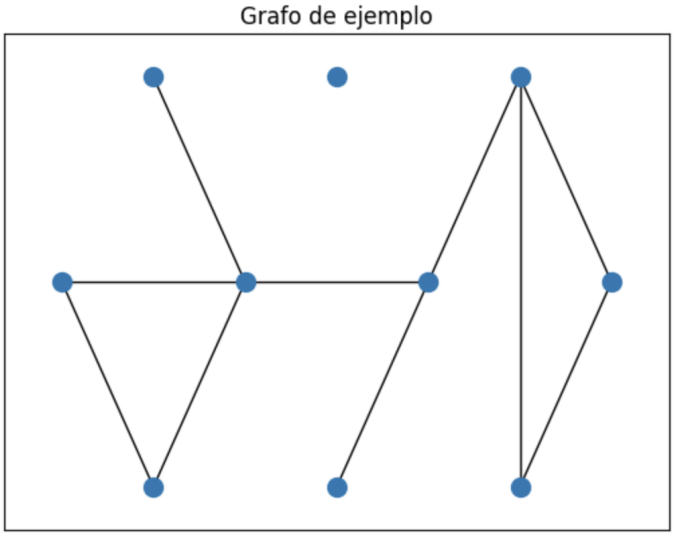

Figura 1: Grafo de ejemplo

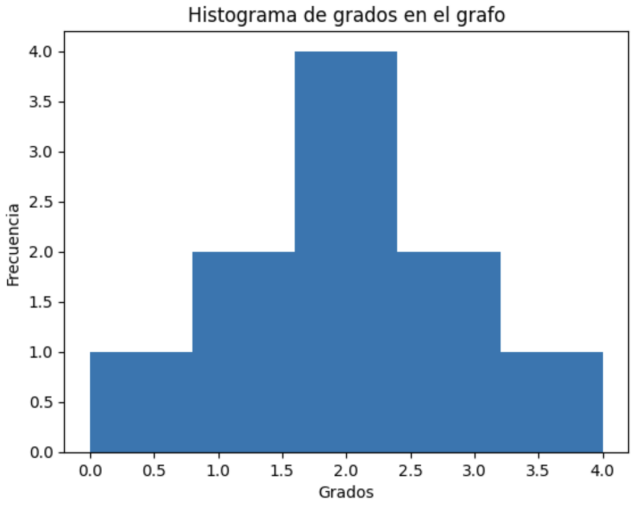

Figura 2: Histograma de grado para el grafo de prueba

Conocer la distribución de los grados de un grafo resulta interesante para comenzar a conocer la red particular con la que se está trabajando.

Power-law

Se dice que la red sigue una distribución de grados "Power-law" si se cumple aproximadamente que $p_k = Ck^{-\alpha}$. Las redes que siguen esta ley tienen un comportamiento de decaimiento lineal, si se grafica la frecuencia de grados en función de los grados, como se

verá a continuación. Muchas de las redes con las que se trabaja en la práctica siguen esta distribución, como por ejemplo las redes sociales, la topología de la red de internet, las redes de potencia, etc. Es decir, estas redes presentan altas frecuencias de baja cantidad de nodos, y a medida que crece la cantidad de nodos, la frecuencia decae exponencialmente.

Para este laboratorio se trabaja con una red de citaciones entre papers de machine learning extraída del paper " Automating the Construction of Internet Portals with Machine Learning" [1]. En la Figura 3 se puede observar un dibujo del grafo.

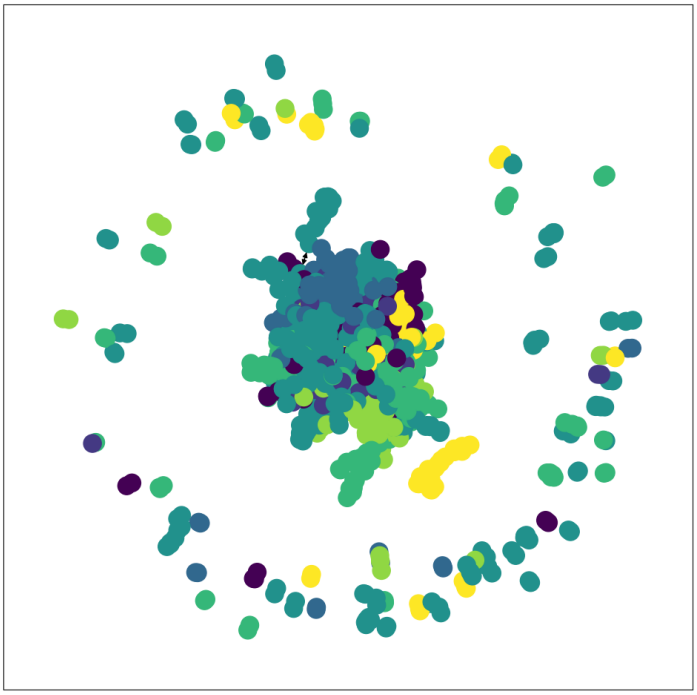

Figura 3: Dibujo del grafo para la red de citaciones de papers.

El grafo presenta 2708 nodos y 10556 aristas no dirigidas. Por otro lado, no existen nodos aislados (es decir, que tengan grado 0), ni tampoco self-loops (no se citan a ellos mismos). Además, existen 7 tipos de clases de nodos en esta red, clasificados según el tópico de cada paper.

Para ver si la red sigue una distribución power-law, lo primero que se hizo fue graficar la distribución de grados de los nodos. En la Figura 4 se puede observar esto mismo, en donde se utilizó escala logarítmica para intentar visualizar correctamente el comportamiento. Además, se realizó en la misma gráfica un fit de una función lineal sobre la escala logarítmica, para obtener una recta que pase por la gran mayoría de puntos. Como se puede observar, existen varios outliers (los nodos con grado muy bajo y muy alto), comportamiento normal de este tipo de redes. Por esto mismo, se realizó un nuevo fit lineal, pero dejando afuera estos puntos, para obtener una recta que describa mejor los grados intermedios.

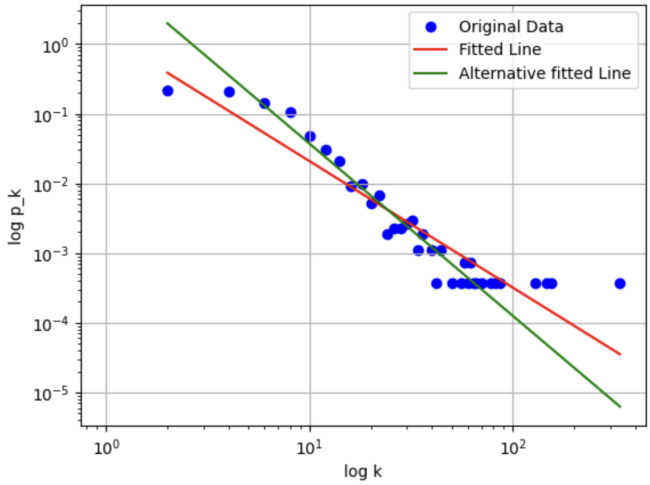

Figura 4: Gráfica para ver comportamiento de ley de potencias.

En redes reales, raramente la distribución de grados sigue una ley de potencias en todo el rango de grados k; generalmente se observa una relación de ese tipo para valores grandes de k, es decir, en la cola de la distribución. En general, cuando se dice que una red tiene power-law, quiere decir que la distribución de grados obedece una ley de potencias para valores grandes de k. En ese caso, se dice que la red es *scale-free* (o libre de escala).

En el histograma de las distribuciones de grados de la red de papers (Figura 5 se debería ver que la distribución de grados tiene fluctuaciones muy grandes para valores altos de k, lo que no permite evaluar claramente la existencia de una ley de potencias. Eso se debe a que, a medida que aumenta el grado k, existen menos nodos que tienen ese grado. Por lo tanto, en cada bin del histograma de la secuencia de grados hay pocas muestras, causando el ruido que debería ver en la cola del gráfico anterior.

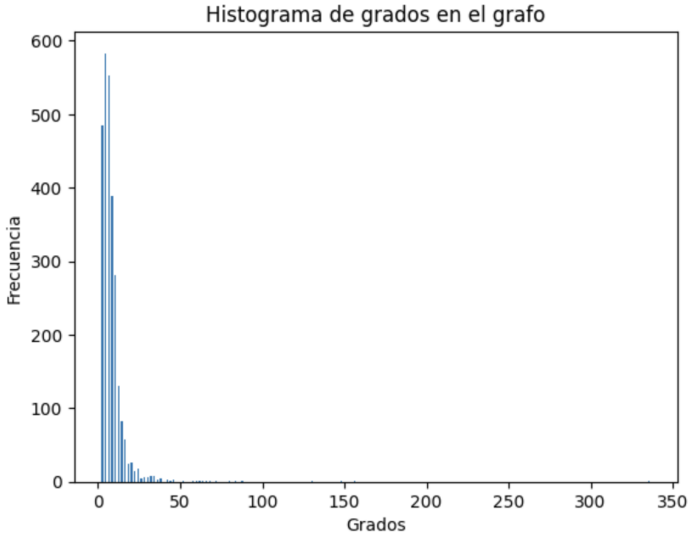

Figura 5: Histograma de grado de nodos para la red de citaciones de papers

Una técnica para que el histograma sea más expresivo consiste en agrandar los bins en la cola del histograma, para que en cada uno hayan más muestras. Por ejemplo, en lugar de usar bins linealmente equispaciados, se utiliza escala logarítmica, como se puede observar en la Figura 6.

Figura 6: Histograma de grado de nodos para la red de citaciones de papers con bins equispaciados en escala logarítmica.

En la Figura 7 se grafica lo mismo que en el histograma, y lo que se puede observar es que los puntos tienden a estabilizarse más hacia una recta. Aunque el primer punto sigue desviando la recta, ya no se tiene la cola pesada del final. En el grafo anterior no se aprecia claramente que siga la ley de potencia debido a que los nodos de grado alto movían la balanza. A partir de esta estrategia, tomando bins de tamaño creciente, se logra que los nodos de grado alto se acumulen en menos puntos al final. Nuevamente se optó por realizar el fit lineal, incluyendo y excluyendo el primer bin.

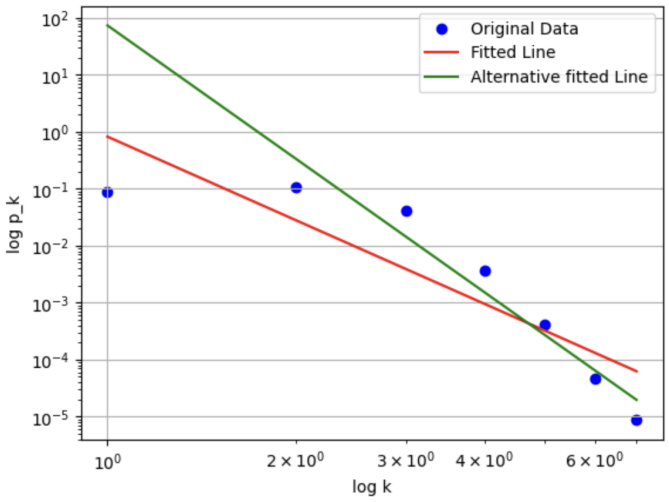

Figura 7: Gráfica de frecuencia de grados tomando bins equispaciados logarítmicamente.

Distribución de Pareto

Con la elección del histograma utilizado, se debería observar una distribución que parece obedecer a una ley de potencias, al menos para valores de k mayores que cierto valor $k_{\min}$, como fue mencionado anteriormente. Para caracterizar este comportamiento, se utiliza la distribución de Pareto: una variable aleatoria X se dice que tiene distribución de Pareto si tiene una función de densidad de probabilidad dada por:

$$p(k) = \begin{cases} Ck^{-\alpha} & \text{si } k \geq k_{\min} \\ 0 & \text{en otro caso} \end{cases}$$

Para la expresión sea efectivamente una densidad, la misma debería sumar 1 con k- ¿infinito. Es decir:

$$\sum_{k=k_{min}}^{\infty} Ck^{-\alpha} = 1$$

$$C = \left(\sum_{k=k_{min}}^{\infty} k^{-\alpha} \right)^{-1}$$

Teniendo n observaciones independientes $k_1, \ldots, k_n$ que provienen de una distribución de Pareto, el estimador de máxima verosimilitud para α es:

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^{n} \log \left(\frac{k_i}{k_{\min}} \right) \right]^{-1}.$$

Por esto mismo se realizó una función que permite estimar α para un grafo, la cual se puede ver en el archivo .ipynb. Luego de evaluar la función para el grafo de citación de papers, el valor obtenido fue $\alpha = 4{,}03$

Assortative mixing

Assortative mixing es la tendencia que tienen los nodos a relacionarse con nodos "similares". Habría que definir qué significa que nodos sean similares. Una posible medida podría ser mirando la cantidad de conexiones que hay entre nodos de una misma clase. Suponiendo que tenemos una red con nodos etiquetados según distintas clases $c_i \in \{1, \ldots, n_c\}$, se define la modularidad como:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

con m la cantidad de aristas de la red, A_{ij} las entradas de la matriz de adyacencia, k_i el grado asociado al nodo i, y $\delta(c_i, c_j)$ que es 0 si $i \neq j$. La modularidad mide la fuerza de conexión entre los grupos de una red. Las redes con alta modularidad tienen conexiones densas entre los nodos dentro de los clusters, pero conexiones escasas entre los nodos en diferentes módulos. Es estricamente menor que 1, y toma valores positivos si hay más aristas entre nodos del mismo tipo que las esperadas, o negativos en caso contrario.

En este caso se trabaja con un dataset de vuelos entre aeropuertos de Estados Unidos, extraído del paper "struc2vec: Learning Node Representations from Structural Identity" [2]. En una estructura de grafos, los nodos son los aeropuertos y las aristas (dirigidas) representan el vuelo que interconecta ambos destinos. A su vez, los nodos se dividen en 4 clases según el nivel de actividad de cada aeropuerto. En la Figura 8 se puede observar una representación del grafo.

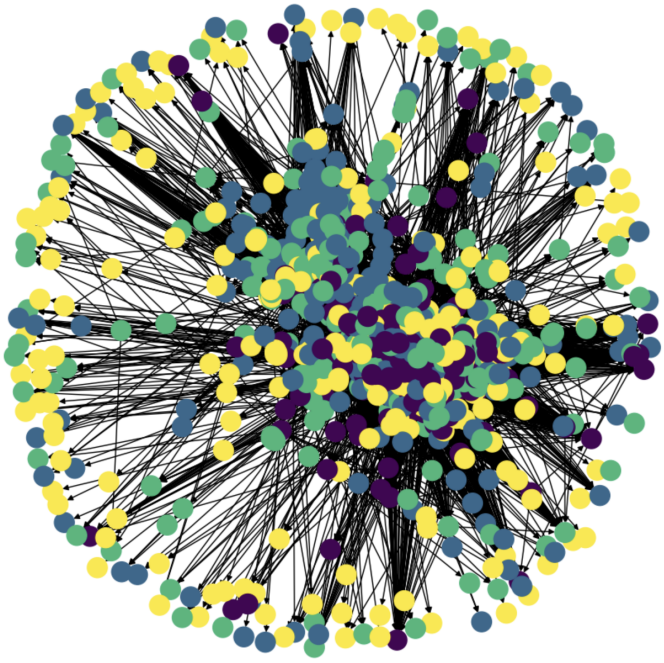

Figura 8: Respresentación del grafo de vuelos entre aeropuertos.

Extrayendo un poco de información del grafo, se obtiene que la cantidad de nodos es 1190, la cantidad de aristas es 13599 (son dirigidas), no existen nodos aislados, claramente no hay self-loops (no tendría sentido que los haya).

La modularidad hayada para este grafo es 0,11, mientras que la modularidad hallada para el grafo de citas entre papers es 0,64.

Como se dijo anteriormente, la modularidad es una medida de qué tanto las clases parecidas se conectan entre sí. Se puede ver que en ambos casos hay más aristas entre nodos del mismo tipo que las esperadas (ambos son positivos), pero en el caso de las citas de papers es bastante mayor, indicando que el grafo tiende más a conectar nodos del mismo tipo que el de los aeropuertos. Esto es perfectamente entendible en el contexto. Es decir, los aeropuertos no seleccionan sus vuelos según la cantidad de vuelos (que es lo que define la clase), sino que se da de manera más jerárquica; hay muchos vuelos de aeropuertos chicos a grandes, mientras que probablemente haya muy pocos entre aeropuertos chicos. El hecho de que probablemente sí haya bastantes vuelos entre aeropuertos grandes, y justamente como son grandes aportan muchos vuelos (aristas), justifica que la modularidad no sea negativa. Por el contrario, es de esperar que los papers con temas en común tiendan a citarse más entre sí, concluyendo en un grafo con alta modularidad.

Detección de dos comunidades

La idea en esta parte es detectar dos comunidades en la base de datos de Zachary's karate club, aplicando diferentes algoritmos. En la Figura 9 se puede observar el grafo del club de Karate, separado según ambas clases. Esta separación será el ground truth contra el cual comparar los algoritmos desarrollados.

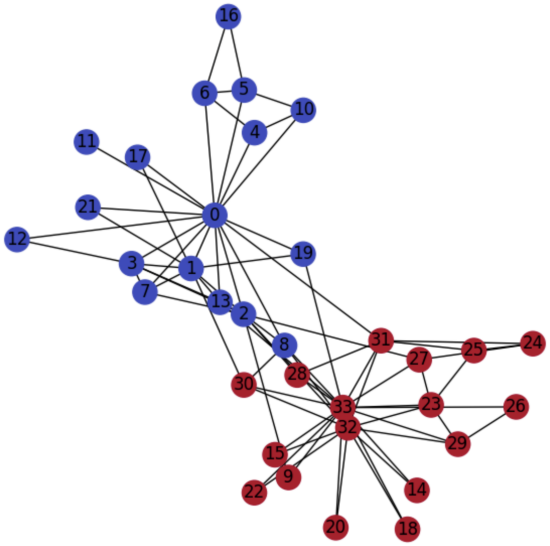

Figura 9: Representación del grafo Zachary's karate club.

Spectral partitioning

Un primer algoritmo de detección de comunidades es spectral partitioning. Presenta la particularidad de que debe conocerse de antemano la cantidad de nodos de cada clase, lo cual en la práctica lo hace menos útil. Lo que hace es generar particiones a partir del Laplaciano L, el cual tiene entradas $L_{ij} = D_{ij} - A_{ij}$. En el archivo .ipynb se puede observar la implementación de este algoritmo. A continuación se probó particionar el grafo de interés; el resultado se puede ver en la **??**. Como se puede observar, se obtiene una partición bastante similar a la original.

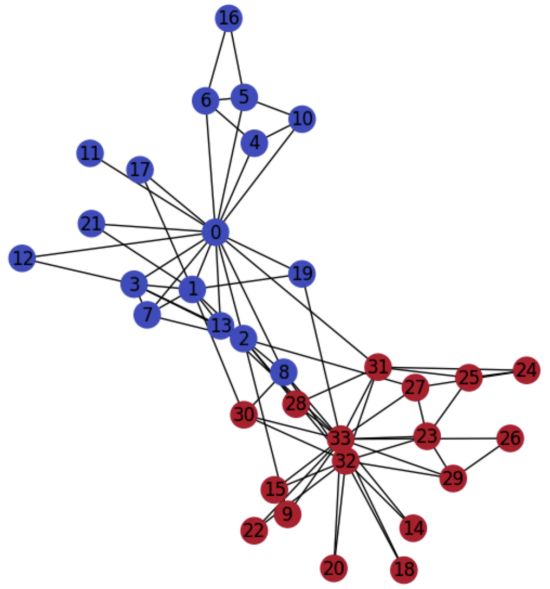

Figura 10: Partición utilizando el algoritmo de spectral partitioning.

A modo de cuantificar los resultados, se introduce el índice de Rand, que cuanto más grande es, quiere decir que las particiones más se parecen. Tiene como valor máximo 1. Se

evaluó para este caso, y el valor obtenido fue 1, por lo cual la partición obtenida es perfecta. Otra posible métrica para ver qué tan buena es la asignación es el índice de Fowlkes-Mallow, que se definde como la media geométrica entre la precision y el recall. Toma valores entre 0 y 1, tomando valor 1 nuevamente cuando las particiones son iguales. Para este caso en particular, el índice de Fowlkes-Mallow vale 1, indicando otra vez que la partición es perfecta.

También se calcularon ambos índices para asignaciones de aristas al azar del grafo. Ahora el índice de Rand vale exactamente 0, mientras que el índice de Fowlkes-Mallow toma valor 0.69.

Spectral modularity maximization

Como fue mencionado anteriormente, el algoritmo de spectral partitioning presenta la desventaja de que debe conocerse la cantidad de nodos pertenecientes a cada comunidad. Por esto mismo se introduce el algoritmo llamado Spectral Modularity Maximization, el cual permite particionar dos comunidades, sin conocer previamente el tamaño de cada una de ellas. La idea principal es encontrar la partición que maximice la modularidad, a partir de su vector propio dominante u_1. La matriz modularidad B tiene entradas $B_{ij} = A_{ij} - \frac{d_i d_j}{2N_e}$. En el archivo .ipynb se puede observar una posible implementación del mismo. Se testeó este algoritmo en el mismo grafo, y la partición obtenida puede verse en la Figura 11.

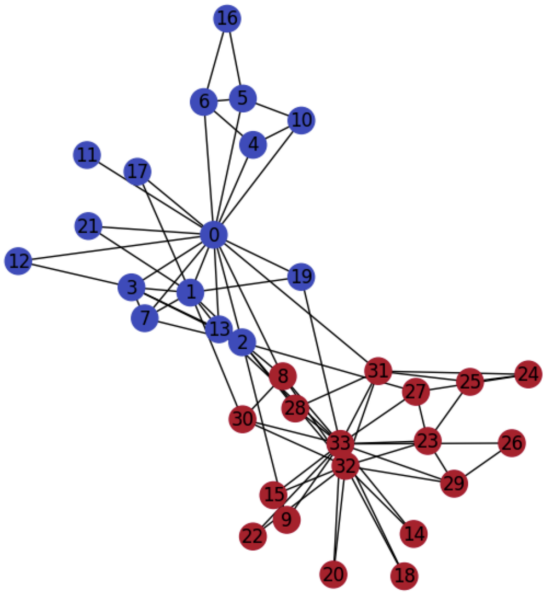

Figura 11: Partición generada con el algoritmo de Spectral Modularity Maximization.

En este caso, las métricas introducidas anteriormente presentan distintos resultados que el algoritmo Spectral Partitioning. El índice de Rand toma valor 0,88, mientras que el índice de Fowlkes-Mallow vale 0,94. Esto se debe a que uno de los nodos (el nodo 8 precisamente) está siendo mal clasificado. Parece ser que es el precio a pagar por no conocer previamente el tamaño de ambas comunidades. Esto puede deberse a que este nodo es un nodo "dudoso", que al conocer previamente los tamaños de las comunidades, termina siendo clasificado para una comunidad, mientras que al no conocer los tamaños previamente, termina siendo asignado a la otra.

A continuación se trabaja con un último dataset real, el cual se trata de datos de blogs sobre política estadounidense, en un momento cercano a la elección del 2004 en ese país. Los

datos son extraídos del paper "The Political Blogosphere and the 2004 US Election: Divided they Blog" [3].

Cada nodo del grafo representa un blog, y existe una arista del nodo i al nodo j si en el blog i hay un link al blog j. Las aristas son dirigidas, aunque para simplificar se lo convierte a no dirigido. Cada blog está etiquetado según su afiliación a uno de los dos grandes partidos de la política estadounidense. Por lo tanto, el grafo tiene únicamente 2 clases.

Extrayendo un poco de información del grafo, se tiene un total de 1490 nodos, 19025 aristas (no dirigidas), existen nodos aislados y existen self-loops (links a su propio blog). En la Figura 12 se puede ver una representación del grafo.

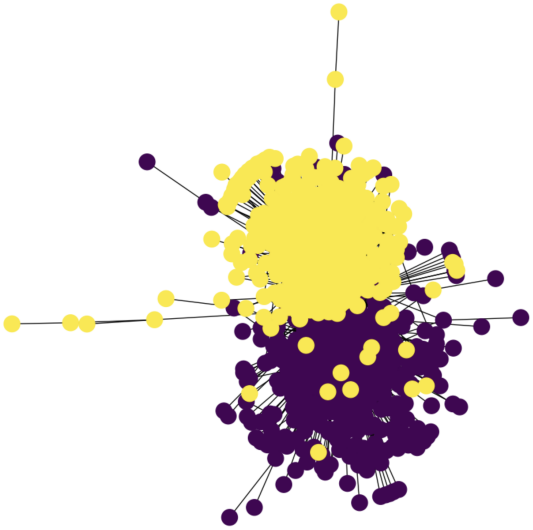

Figura 12: Respresentación del grafo de links entre blogs sobre política.

Como fue mencionado anteriormente, en datasets reales suele ser imposible conocer previamente la cantidad de nodos de cada clase. Por lo tanto se utilizará el algoritmo de Spectral Modularity Maximization para generar ambas comunidades. En la Figura 13 se puede observar la partición generada. A priori parece ser una partición que tiene sentido, aunque hay nodos que claramente están siendo mal clasificados. De todas maneras, visualmente resulta muy difícil realizar un análisis, por lo cual se cuantifican los resultados según las métricas.

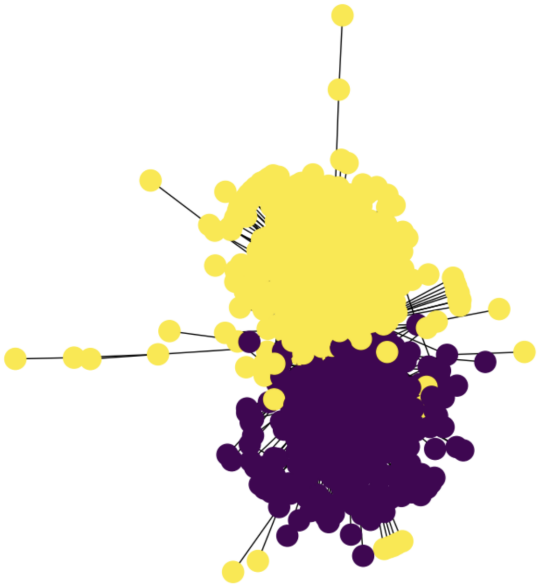

Figura 13: Asignación de comunidades según algoritmo Spectral Modularity Maximization.

El índice de Rand obtenido en este caso es 0,78, mientras que el índice de Fowlkes-Mallow vale 0,89. Quiere decir que la partición que se está logrando tiene sentido, aunque claramente no es perfecta. Analizando cualitativamente ambas figuras (los ground truth vs la partición generada), se puede ver que hay diferencias entre ambas clasificaciones de nodos. Se nota claramente que algunos nodos que eras amarillos se clasifican como violetas y viceversa. Podría suceder que haya nodos muy difíciles de clasificar porque sean bastante anómalos. Por ejemplo, algún blog con afiliación hacia un político que cite mucho blogs del otro político para hacer una crítica o algo del estilo; también podría haber un blog bastante neutro que funcione de "puente" entre comunidades y sea difícil de clasificar.

Además, este algoritmo presenta problemas de resolución. Esto quiere decir que tiene dificultad para encontrar comunidades pequeñas, dentro de comunidades más grandes. Aplicado a este caso, puede estar asignando subgrupos más conectados entre sí a la otra comunidad, teniendo un alto impacto sobre el desempeño.

De todas formas, el hecho de que la mala clasificación corra para ambas afiliaciones políticas lleva a pensar que los errores no necesariamente provengan de la falta de información del tamaño de las componentes, dado que se equivoca para ambos lados y podría suceder lo mismo utilizando el otro método. Por eso, se decidió hacer un poquito de "trampa" y ver cuántos nodos de cada clase hay, para probar el algoritmo de Spectral Partitioning.

La partición generada con este último se puede observar en la Figura 14. Como se puede ver, el problema de la detección no viene solamente de no saber los tamaños de la comunidad. Aunque el método de Spectral Modularity Maximization paga el precio de no saber la cantidad de nodos en cada comunidad, maximizar la modularidad obtiene un mejor resultado que el método de bisección que minimiza el corte. Esto podría ser más influyente que en el grafo del club de Karate, debido al tamaño y complejidad del grafo.

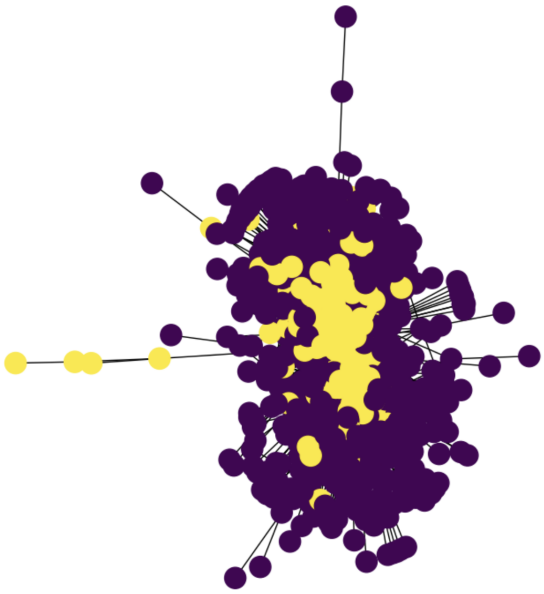

Figura 14: Partición espectral para la base de links sobre blogs de política.

Cuantitativamente se obtuvo un índice de Rand de 0,33, mientras que el índice de Fowlkes-Mallow vale 0,52. Según ambas métricas, efectivamente este algoritmo está asignando clases aleatoriamente a los nodos.

Referencias

[1] https://link.springer.com/article/10.1023/A:1009953814988

[2] https://arxiv.org/pdf/1704.03165.pdf

[3] https://dl.acm.org/doi/10.1145/1134271.1134277