



UNIVERSIDAD DE GRANADA

MINERÍA DE MEDIOS SOCIALES
MÁSTER CIENCIA DE DATOS E INGENIERÍA DE COMPUTADORES

PRÁCTICA 1

ANÁLISIS Y VISUALIZACIÓN CON GEPHI

Autor

Ignacio Vellido Expósito
ignacioove@correo.ugr.es
79056166Z



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍAS INFORMÁTICA Y DE
TELECOMUNICACIÓN

CURSO 2020-2021

Índice

1. Resultados globales	2
2. Análisis de la Red	5
3. Análisis de Centralidad	6
4. Estudio de las Comunidades	13
5. Anexo: Gráficos adicionales	17

1. Resultados globales

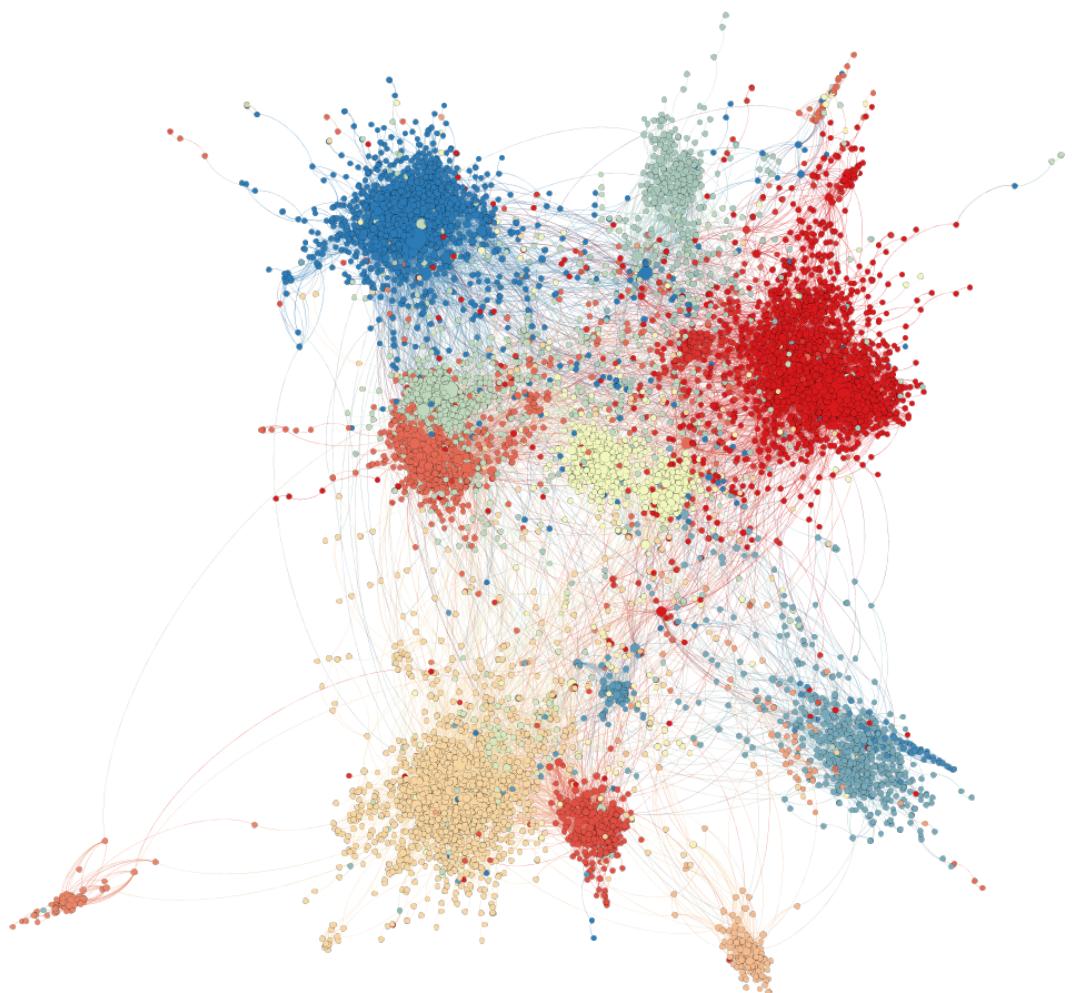


Figura 1: Topología de la red. El color indica el país de cada usuario.

Medida	Valor
Número de nodos N	7,624
Número de enlaces L	27,806
Número máximo de enlaces L_{max}	58117752
Densidad del grafo L/L_{max}	0.001
Grado medio $\langle k \rangle$	7.294
Diámetro d_{max}	15
Distancia media d	5.232237269
Coeficiente medio de clustering $\langle C \rangle$	0.285
Número de componentes conexas	1
Número de nodos componente gigante (y %)	7,624 (100)
Número de aristas componente gigante (y %)	27,806 (100)

Figura 2: Medidas globales de la red.

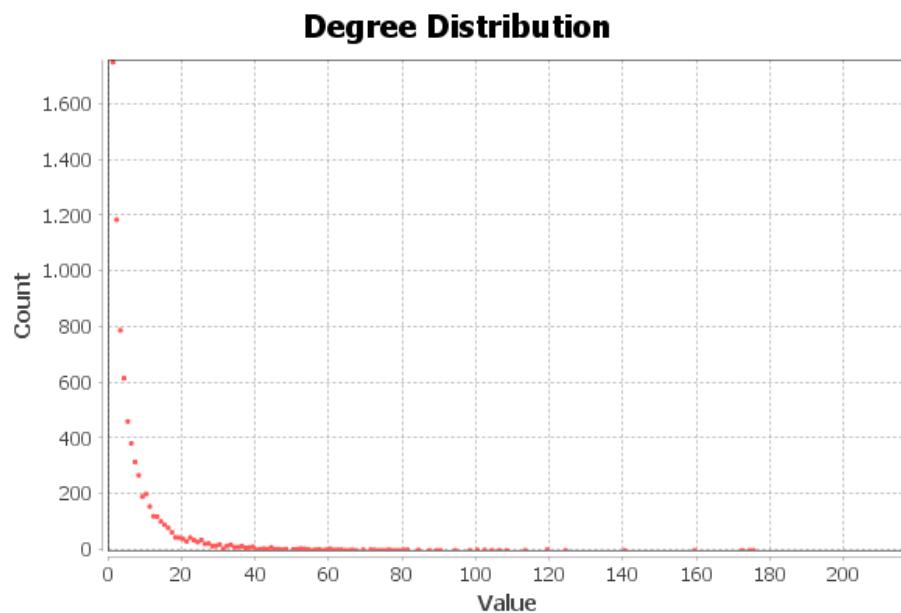


Figura 3: Distribución de grados.

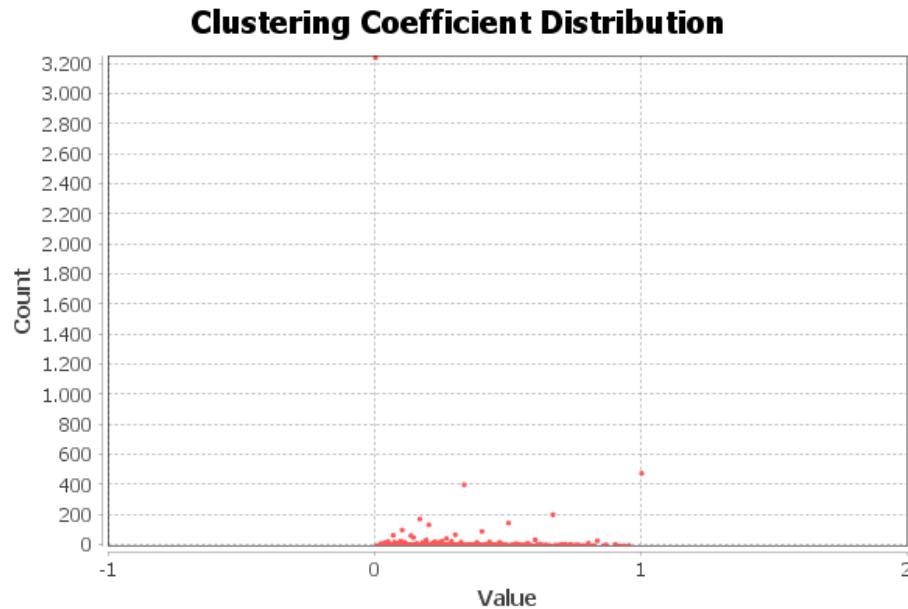


Figura 4: Distribución del clustering medio de la red.

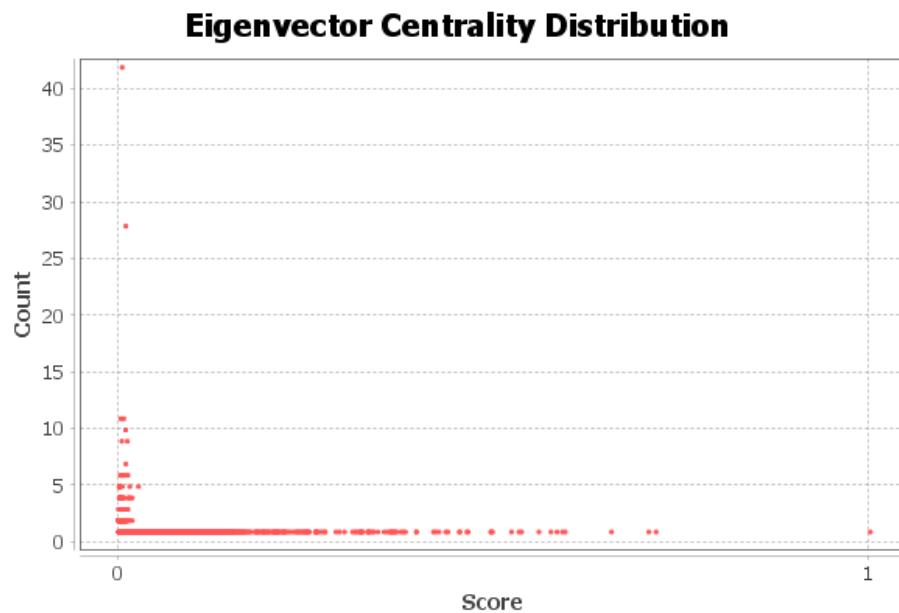


Figura 5: Distribución de los valores de vector propio.

2. Análisis de la Red

Para esta práctica se ha usado la red social **LastFMAAsia**, que codifica conexiones mutuas de amistad entre usuarios asiáticos de la plataforma LastFM. El dataset incluye además una variable *target* que codifica el país de origen de cada uno de los usuarios, con un total de 17 regiones posibles.¹

LastFM es una red social centrada en el ámbito musical, donde los usuarios pueden escuchar, comentar y debatir sobre sus gustos e intereses. De forma similar a Facebook, se permiten crear comunidades y foros de debate donde compartir información.

Tenemos por tanto una red no dirigida y sin pesos, con una única componente conexa. La red se caracteriza por tener una gran dimensión pero muy baja densidad (probablemente típica en redes de este tamaño), aunque la distancia media no es alta debido a un buen grado medio global en la red.

Apreciamos también un buen coeficiente de clustering, pero un tanto bajo para los habituales en redes sociales. Esto se puede notar fácilmente en la Figura 1, donde aunque la mayor parte de los nodos pertenecen a algún hub (ya sea de mayor o menor tamaño) existe un buen número de ellos que se ubican en “zonas de paso” entre países.

En la Figura 4 vemos que la tendencia a este valor un tanto bajo se ve afectada por tener un 40 % de los nodos un coeficiente muy cercano a cero.

Sobre la distribución de grados, contamos con una media de 7.294 y una desviación típica de 11.499. Esta media es un tanto baja en comparación con redes de amistad como Facebook, aunque en este caso al ser un red centrada en música no resulta extraño pues es probable que los nodos no estén conectados con personas que conozcan personalmente.

En la Figura 3 vemos claramente una ley de la potencia en la distribución, indicándonos que tenemos una red libre de escala. Aquí se refleja mejor la alta desviación en el grado de los nodos, pues existe una buena cantidad de ellos con más de 100 enlaces.

Estos actores probablemente correspondan a creadores habituales de contenido (posts, reviews...) y sean muy influyentes en el manejo de información de la red.

¹Referencia: <http://snap.stanford.edu/data/feather-lastfm-social.html>

3. Análisis de Centralidad

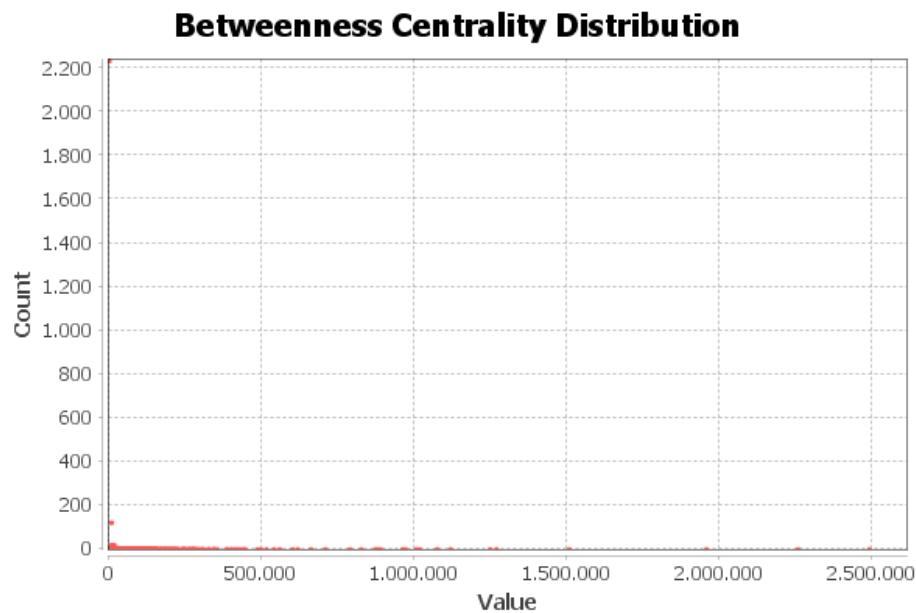


Figura 6: Distribución de valores de intermediación.

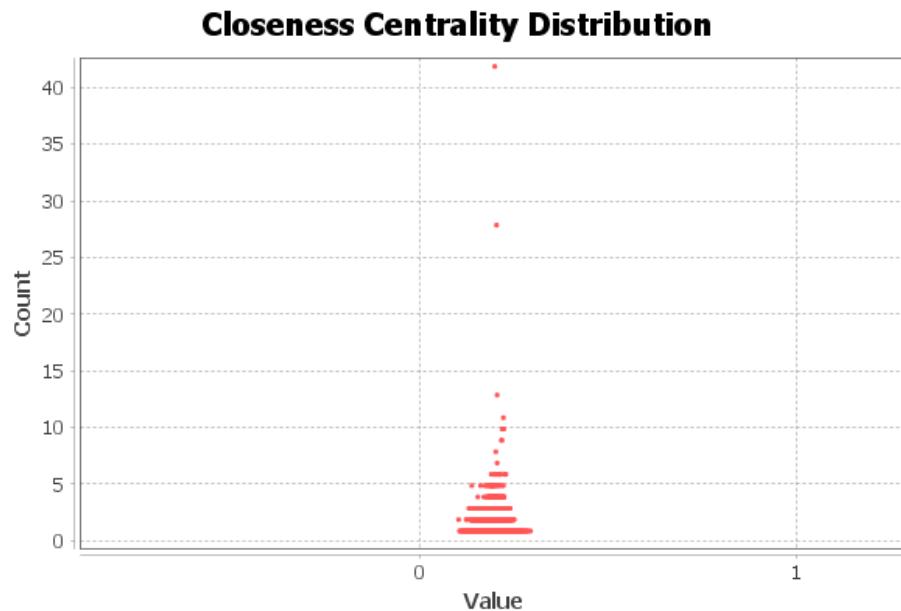


Figura 7: Distribución de valores de cercanía.

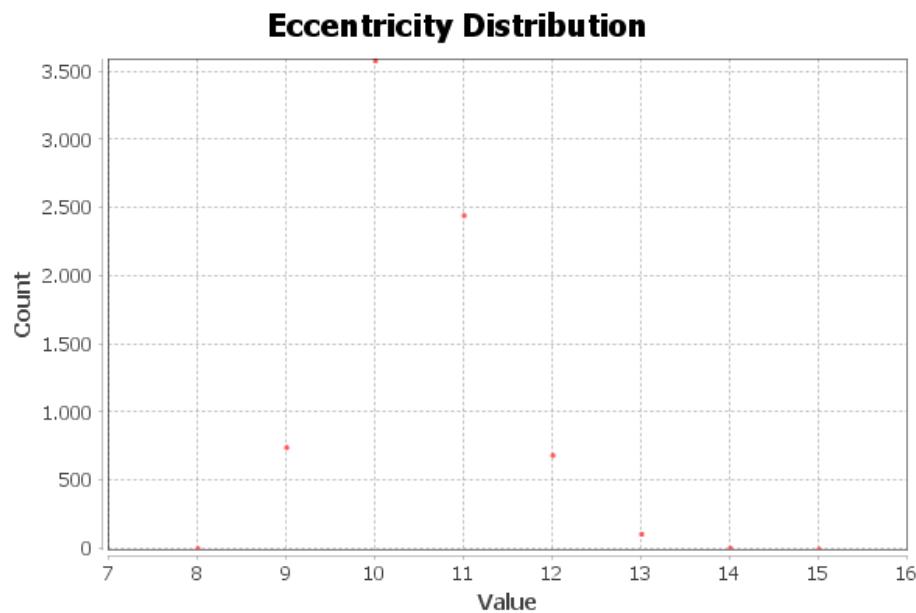


Figura 8: Distribución de valores de excentricidad.

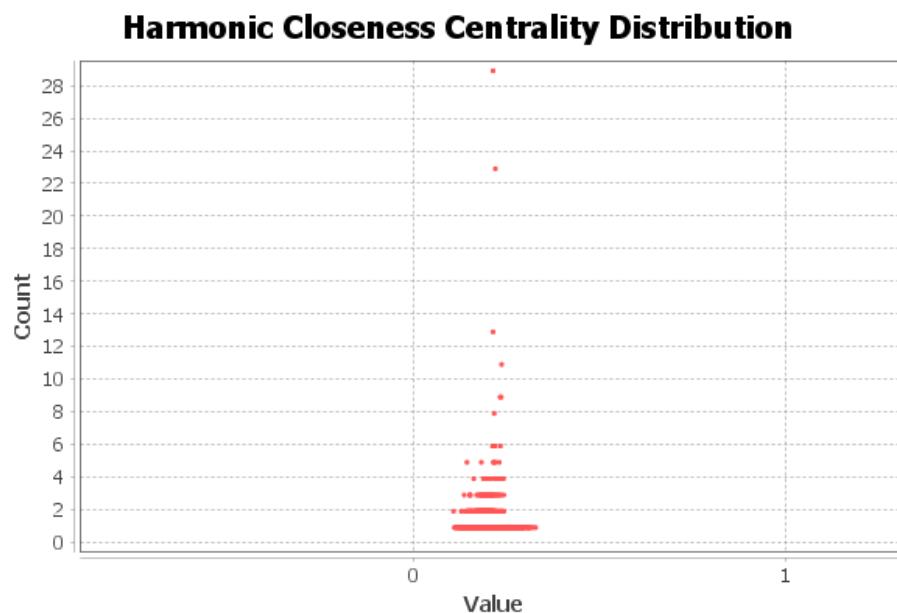


Figura 9: Distribución de valores de cercanía harmónica.

Centralidad de Grado	Intermediación	Cercanía	Vector propio
7237 - 216	7199 - 2,612,617	7199 - 0.2907	7237 - 1.0000
3530 - 175	7237 - 2,486,453	7237 - 0.2856	3240 - 0.7149
4785 - 174	2854 - 2,253,302	4356 - 0.2816	3597 - 0.7052
524 - 172	4356 - 1,953,690	2854 - 0.2803	763 - 0.6555
3450 - 159	6101 - 1,504,994	5454 - 0.2798	2083 - 0.5940

Figura 10: Tabla de actores más relevantes (identificador-valor).

Respecto a la centralidad, como se comentó anteriormente estos outliers es muy probable que correspondan a creadores de contenido, ya que sus valores se encuentran extremadamente separados de la media en la red.

Acerca de la intermediación, vemos en la Figura 6 que la variación en la distribución es enorme, donde la mayoría de nodos se ubican por debajo de los 250,000. Este resultado no es extraño, el grado medio es relativamente bajo en esta red de amistad, por lo que solo un porcentaje pequeño de los nodos hacen de conexiones entre las diferentes comunidades.

Hacemos notar que los nodos con alta intermediación tienen también los valores más altos de cercanía. No nos queda del todo claro cuál puede ser el motivo tras ello, pero es probable que estos nodos estén altamente conectados con dos o más hubs en la red, tal y como muestra la Figura 13.

También nos fijamos en los valores de vector propio tras 100 iteraciones. Es de destacar una alta diferencia entre el primer actor (7237) respecto al resto, además contando con un valor de 1. Esto nos indica que la ubicación en la que se ubica en la red es muy buena, y no resulta extraño el encontrarnos enlaces con 3240, 3597 y 2083, también tres de los actores con mayor valor de vector propio.

En las Figuras 11 y 12 se muestran los vecinos a profundidad uno y dos del nodo 7237. Vemos que ya de por sí las conexiones del actor son buenas, pero además a partir de sus enlaces es capaz de extenderse a la zona más remota de la red (la inferior izquierda de la topología global).

Por último, en la Figura 13 vemos la unión de los vecinos a profundidad dos del par de nodos 7237-7199. Destacamos dos cosas de esta figura: por un lado, la gran importancia de estos actores puesto que con un máximo de dos enlaces alcanzamos un 26.3% de la red (2043 nodos); por otra, la influencia en el flujo de información ya que sabemos que **en media** añadiéndole solo cuatro enlaces más (para un total de seis) podemos transmitir a toda la red.

Estos hechos se hacen notar respectivamente en los altos valores de cercanía (una buena parte de los nodos se conectan a estos dos en pocos saltos) e intermediación (hacen de enlace entre varias comunidades).

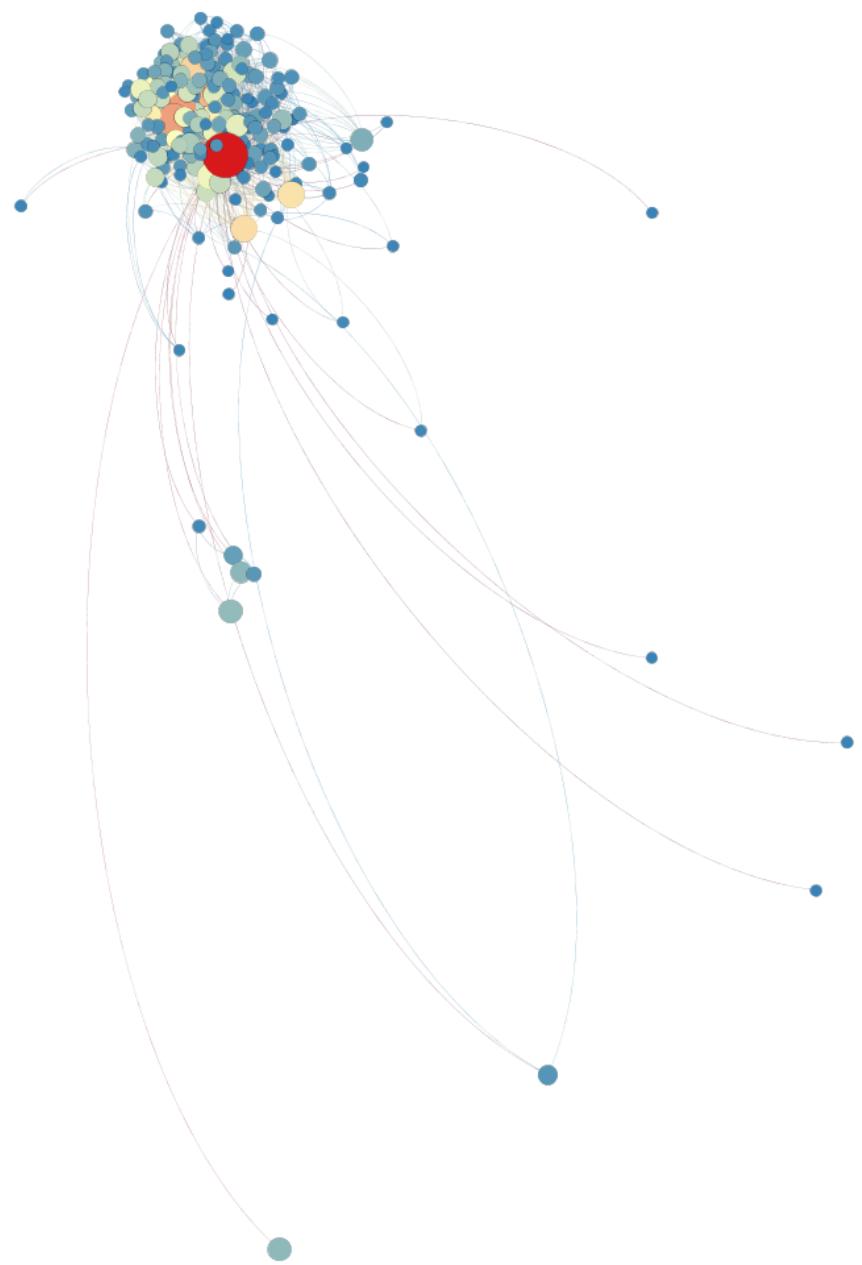


Figura 11: Vecinos del nodo 7237. A mayor grado mayor tamaño, más rojo mayor valor de vector propio.

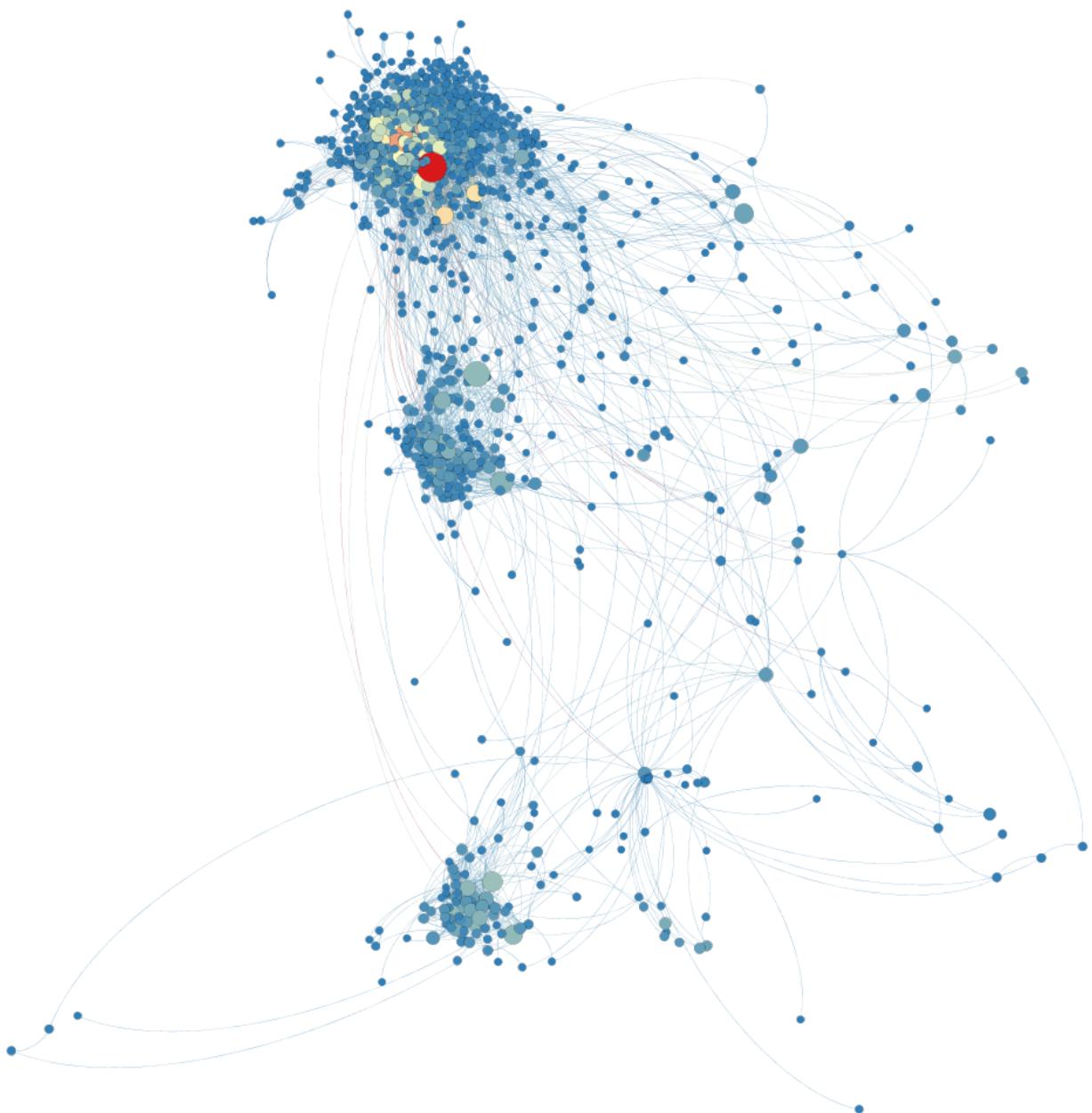


Figura 12: Vecinos del nodo 7237 a profundidad 2.

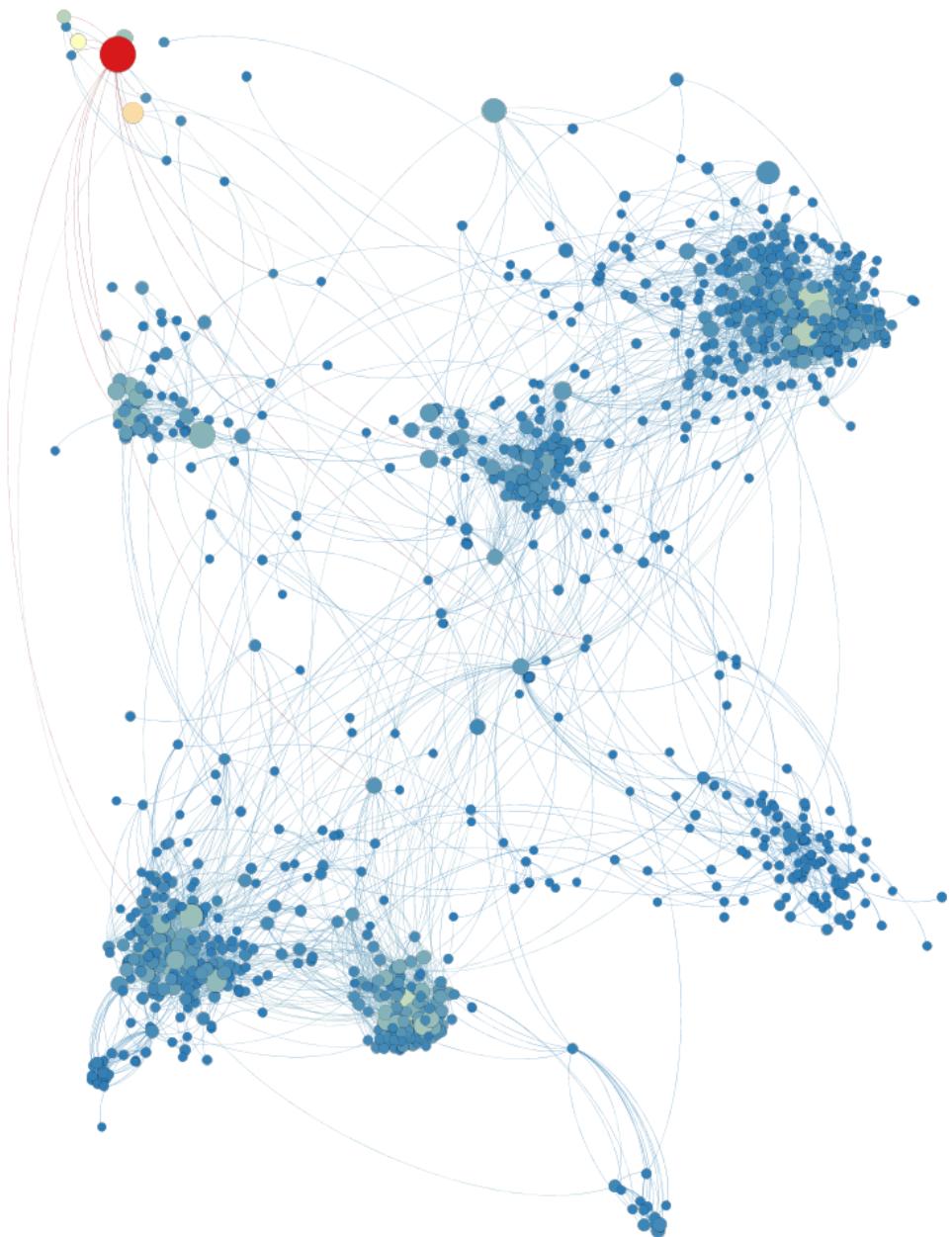


Figura 13: Vecinos del nodo 7199 a profundidad 2.

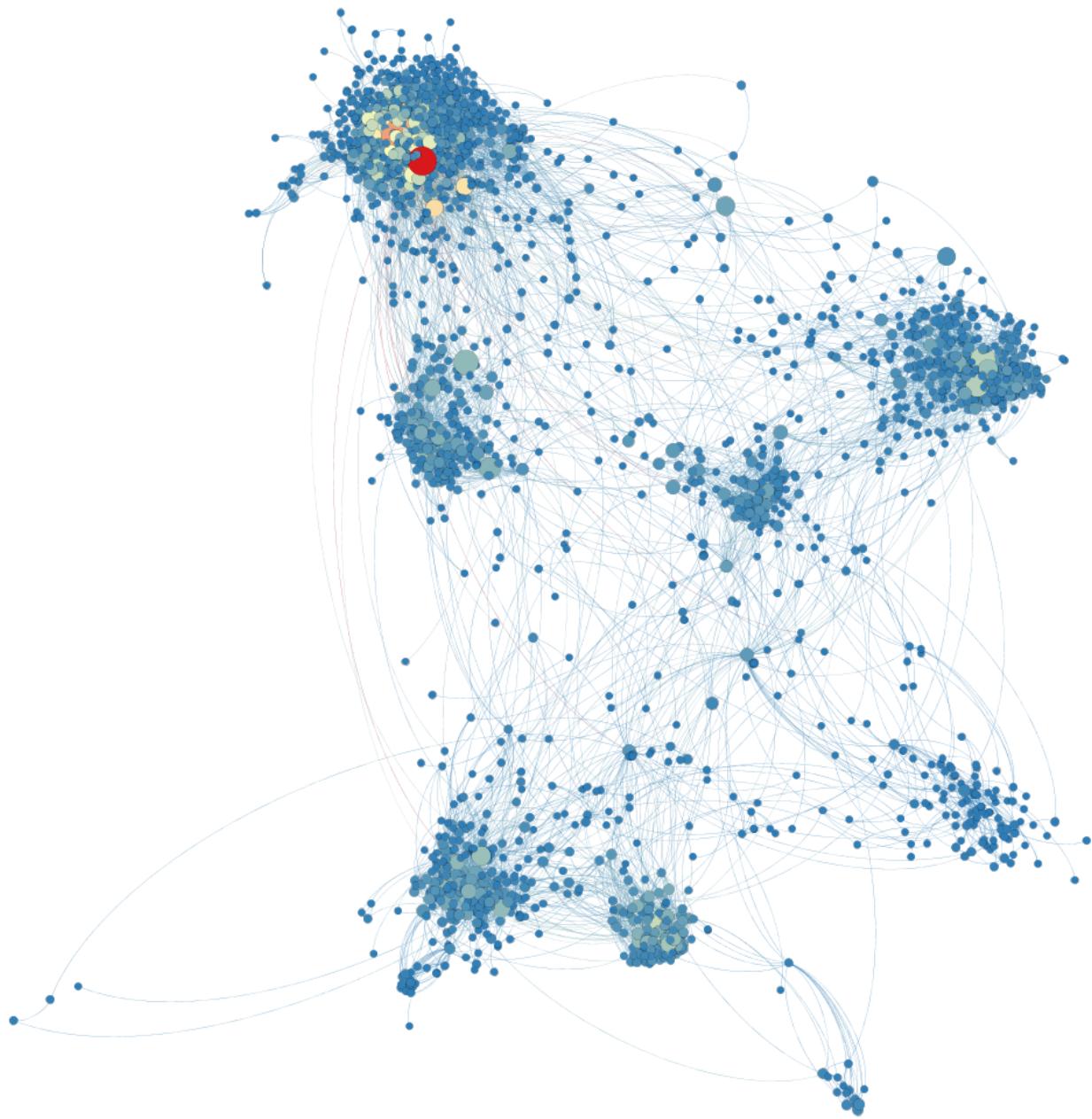


Figura 14: Unión de los vecinos de los nodos 7237 y 7199 a profundidad 2.

4. Estudio de las Comunidades

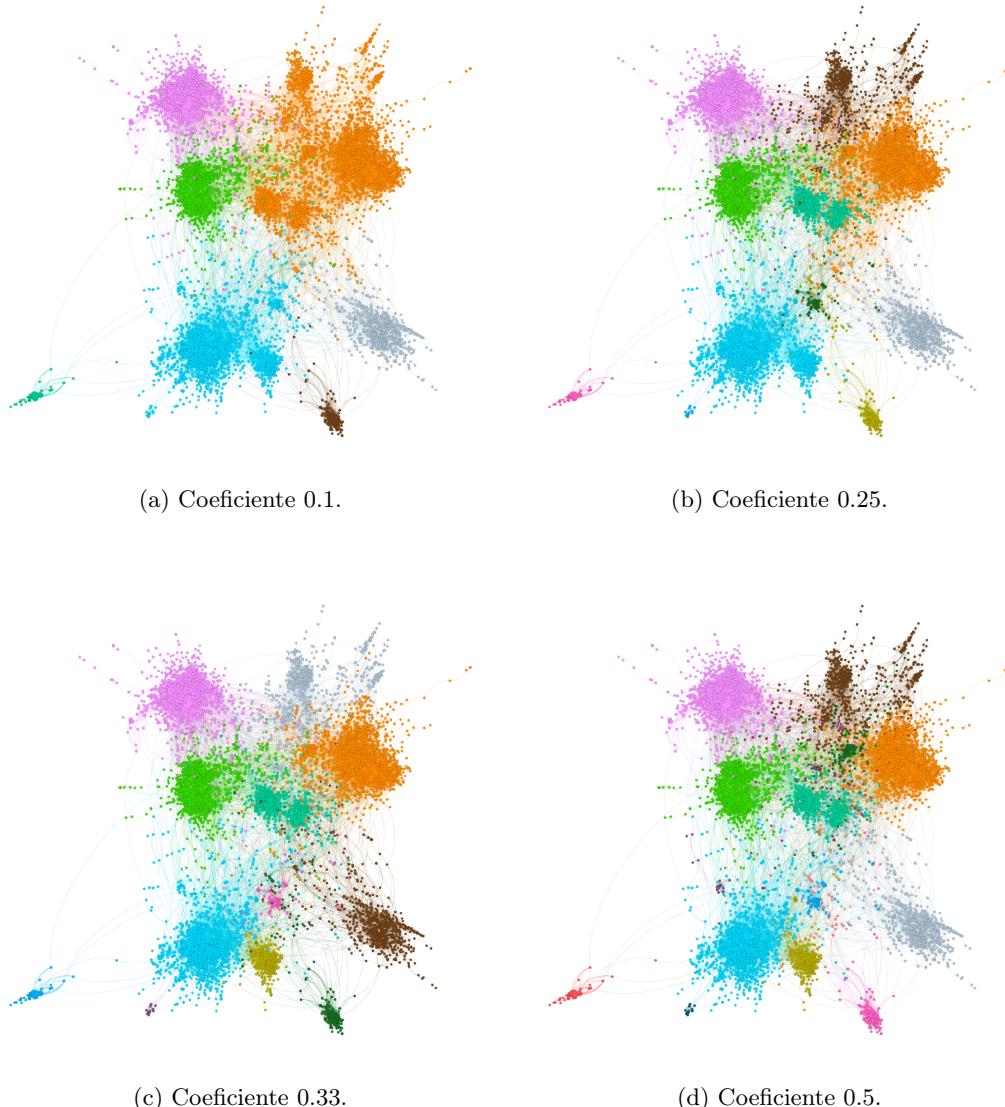
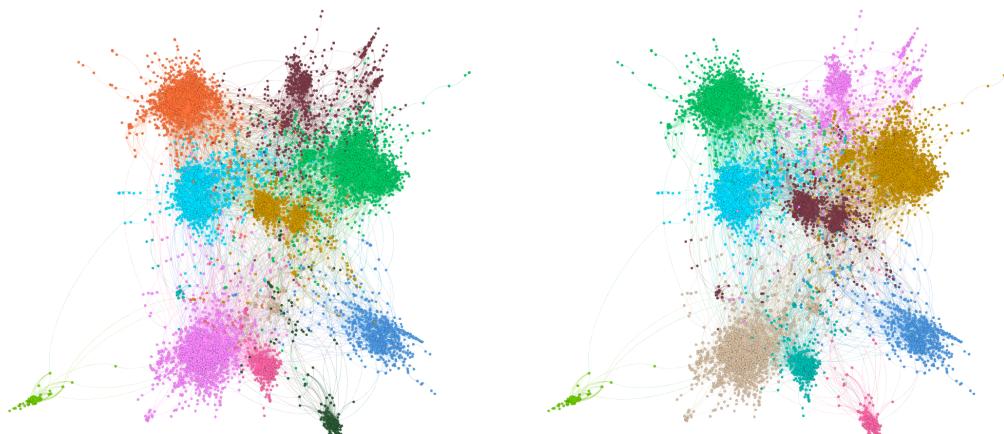


Figura 15: Comunidades detectadas por el algoritmo Leinen para diferentes coeficientes.



(a) Coeficiente 0.5.

(b) Coeficiente 1.



(c) Coeficiente 2.

(d) Coeficiente 3.

Figura 16: Comunidades detectadas por el algoritmo Lovaina para diferentes coeficientes.

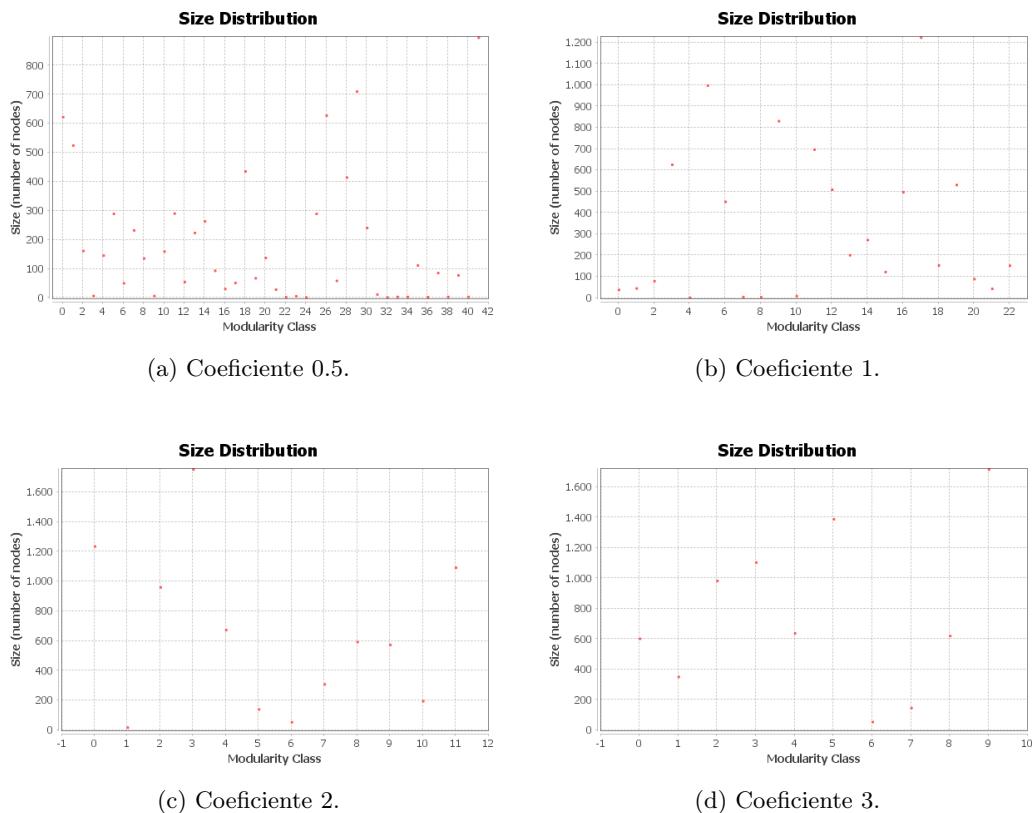


Figura 17: Gráfico de distribuciones para el algoritmo Lovaina en diferentes coeficientes.

Algoritmo	Coeficiente	Modularidad	Clústers
Leinen	0.1	0.941	7
	0.25	0.913	11
	0.33	0.901	11
	0.5	0.878	15
Lovaina	0.5	0.800	42
	1	0.813	23
	2	0.806	12
	3	0.803	10

Figura 18: Modularidad y número de clústers obtenido por cada algoritmo.

La referencia sobre los países puede ser un buen comienzo de partida para la búsqueda de buenas comunidades, pero es importante no dejarse engañar por esto pues existen hubs de diferentes regiones altamente conectados entre sí. Además, algunas de las fronteras entre los hubs de los países son difusas y existen zonas de baja conectividad fuertemente entremezcladas.

El objetivo de la búsqueda de comunidades es después de todo encontrar grupos de usuarios con intereses comunes que nos permitan simplificar el manejo de la red y la transmisión de información. La ubicación regional probablemente sea relevante en la semántica de nuestra red pero no debe ser el factor decisivo para una selección correcta de comunidades.

A partir de la estructura de la red vemos que la zona central y superior del grafo no muestra una estructura modular clara. Por el contrario, la zona inferior muestra más claramente una partición en un mínimo de cuatro o cinco comunidades, y esto es detectado perfectamente por el método Leinen independientemente del coeficiente, pero no tanto por Lovaina, donde es necesario empujar el coeficiente hacia arriba para separar mejor estos hubs.

En cuanto a los efectos en la modularidad y el número de clústers que muestra la Figura 18, nos fijamos en que la calidad de las particiones con Lovaina no varía de manera substancial, aunque sí lo hace el número de clústers que forma.

Sobre las comunidades en sí, visualmente concluimos que un número de diez parece adecuado para esta red, tal y como se consigue en la Figura 16.d. De esta forma contaría con cinco clústers bien diferenciados en la parte inferior conectándose a un núcleo central a su vez dividido en cuatro. Este núcleo contaría con la mayoría de nodos de la red fuertemente interconectados entre sí. Por último, nos quedaría un hub denso en la parte superior izquierda con muchas conexiones al núcleo grande.

En referencia al significado semántico de estas comunidades, es de esperar que los usuarios se agrupen por estilos y géneros de música favoritos donde los intercambios de información vayan sobre esos temas.

Con una partición en diez podríamos suponer que en el núcleo central predominan los géneros de moda en la cultura asiática actual (la red se construyó en 2020), como es el K-Pop, el J-Pop y el C-Pop, y que en las comunidades inferiores predominan estilos relacionados con estos géneros pero de ámbito menos popular, como el rock, el rap o el pop anglosajón.

Por último, tendríamos un hub interesante en la esquina inferior izquierda, conectándose al resto de la red por muy pocos nodos. Este hub nos podría dar información sobre los géneros menos relevantes en la comunidad, como podría ser el metal, el jazz, o la electrónica.

5. Anexo: Gráficos adicionales

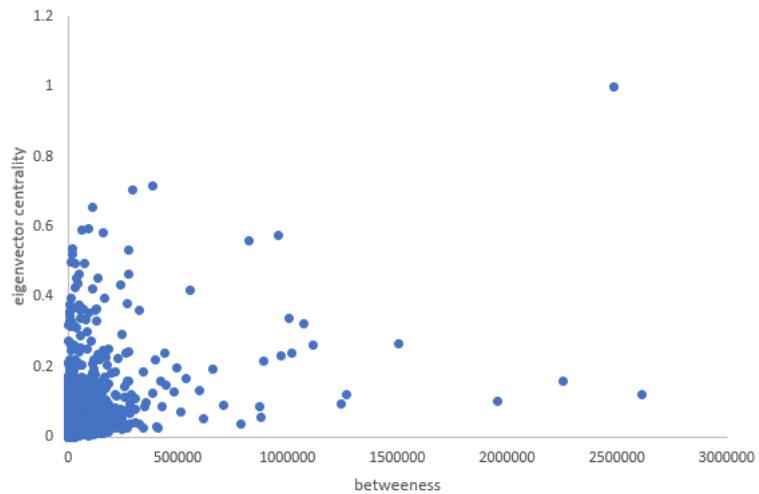


Figura 19: Relación intermediación-vector propio.

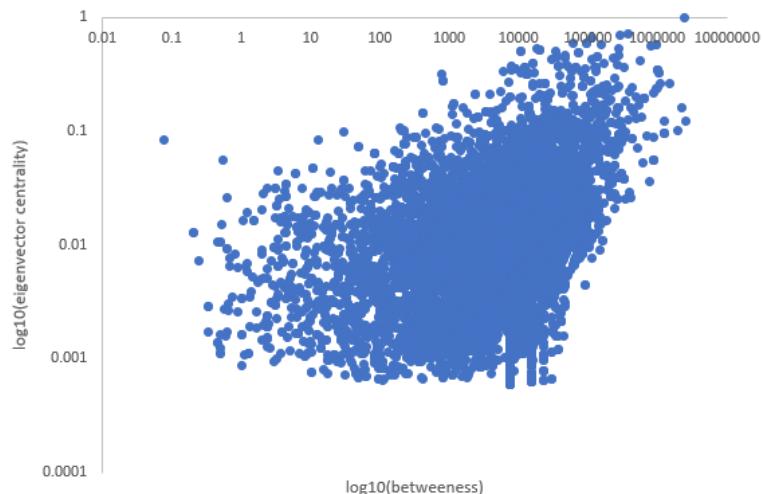


Figura 20: Relación intermediación-vector propio en escala logarítmica.

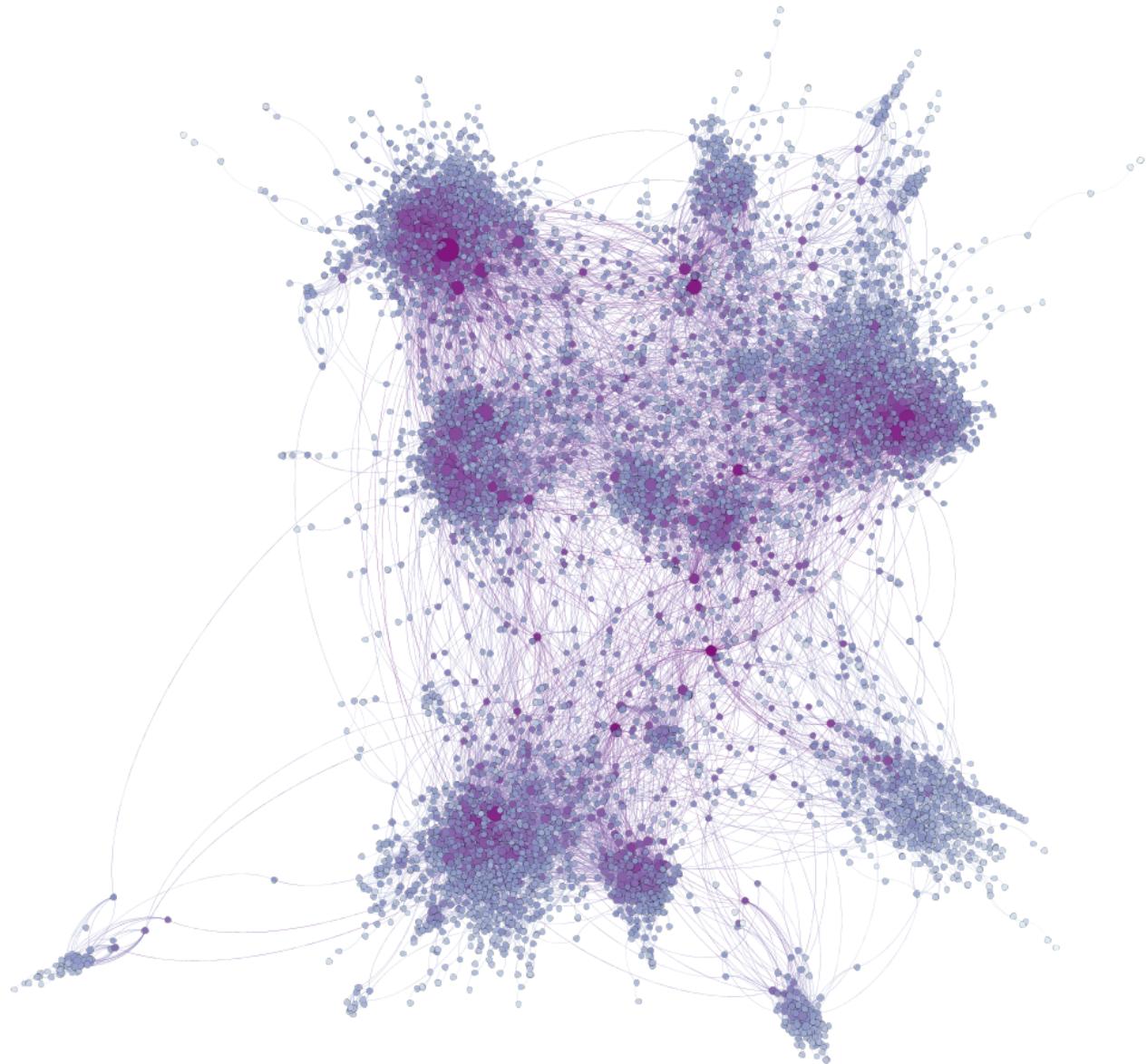


Figura 21: A mayor tamaño de nodo mayor grado. Mayor intensidad de violeta implica mayor cercanía.

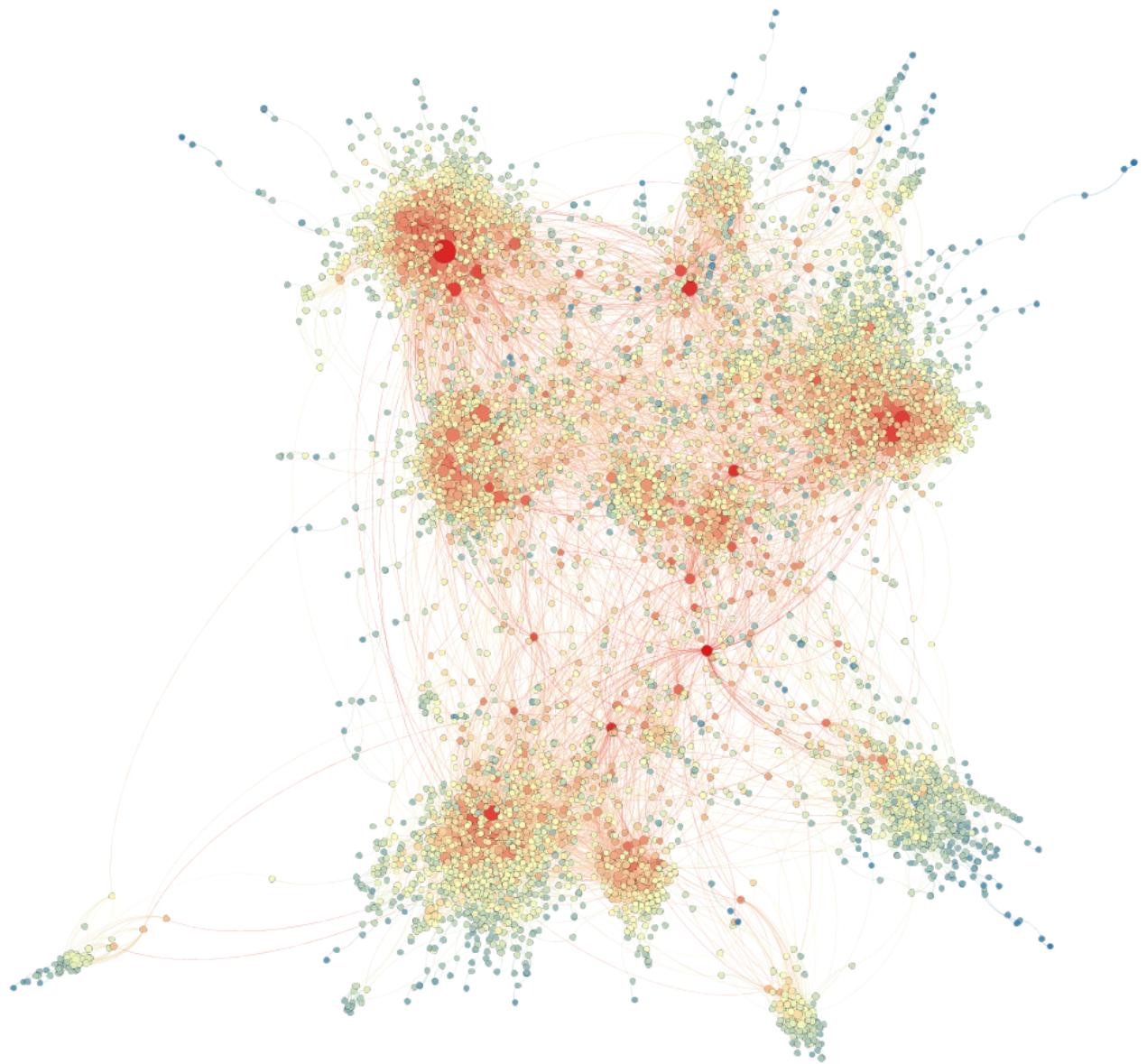


Figura 22: A mayor tamaño de nodo mayor grado. Azul implica menor cercanía, rojo más.

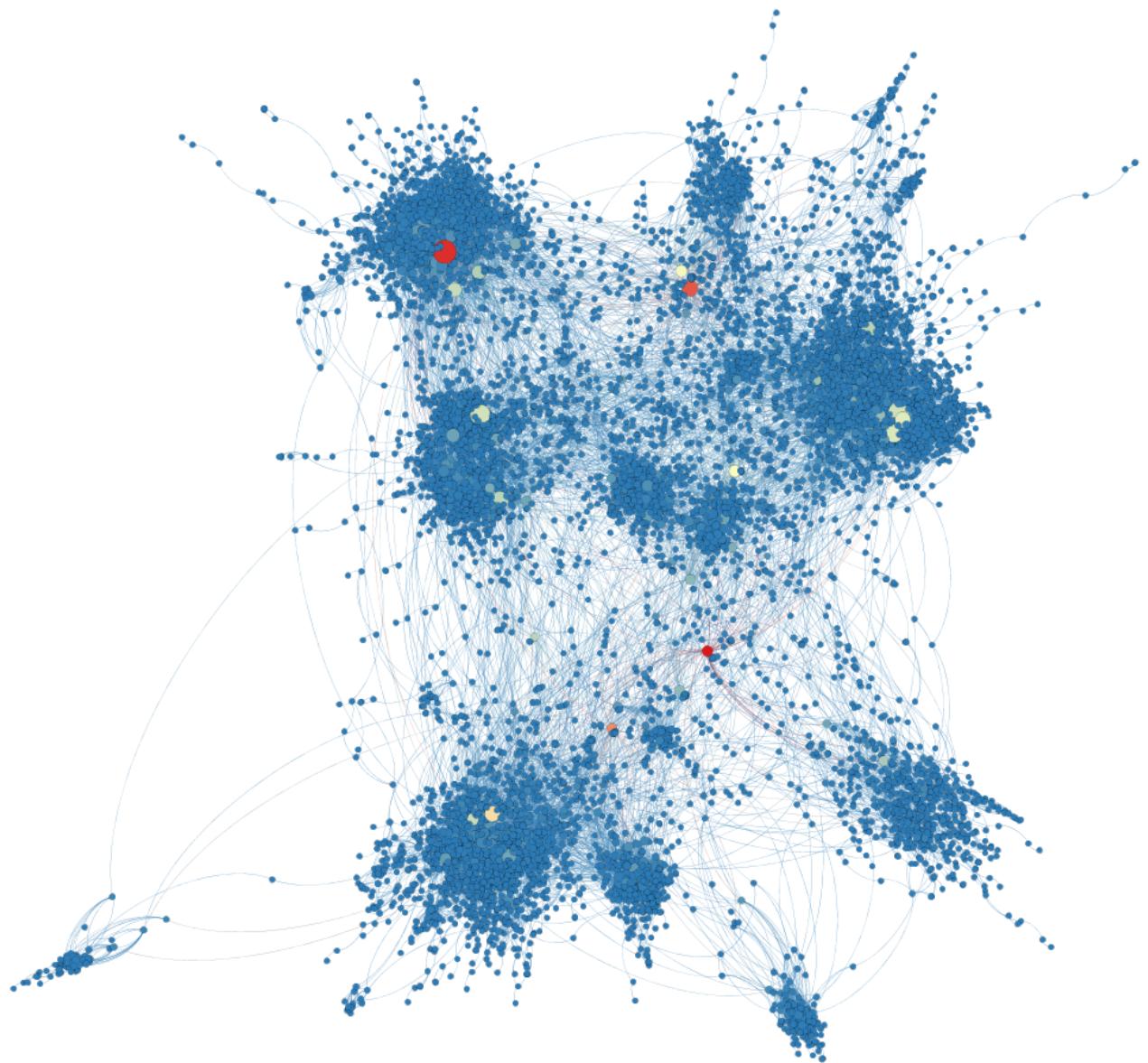


Figura 23: A mayor tamaño de nodo mayor grado. Azul implica menor intermediación, rojo más.

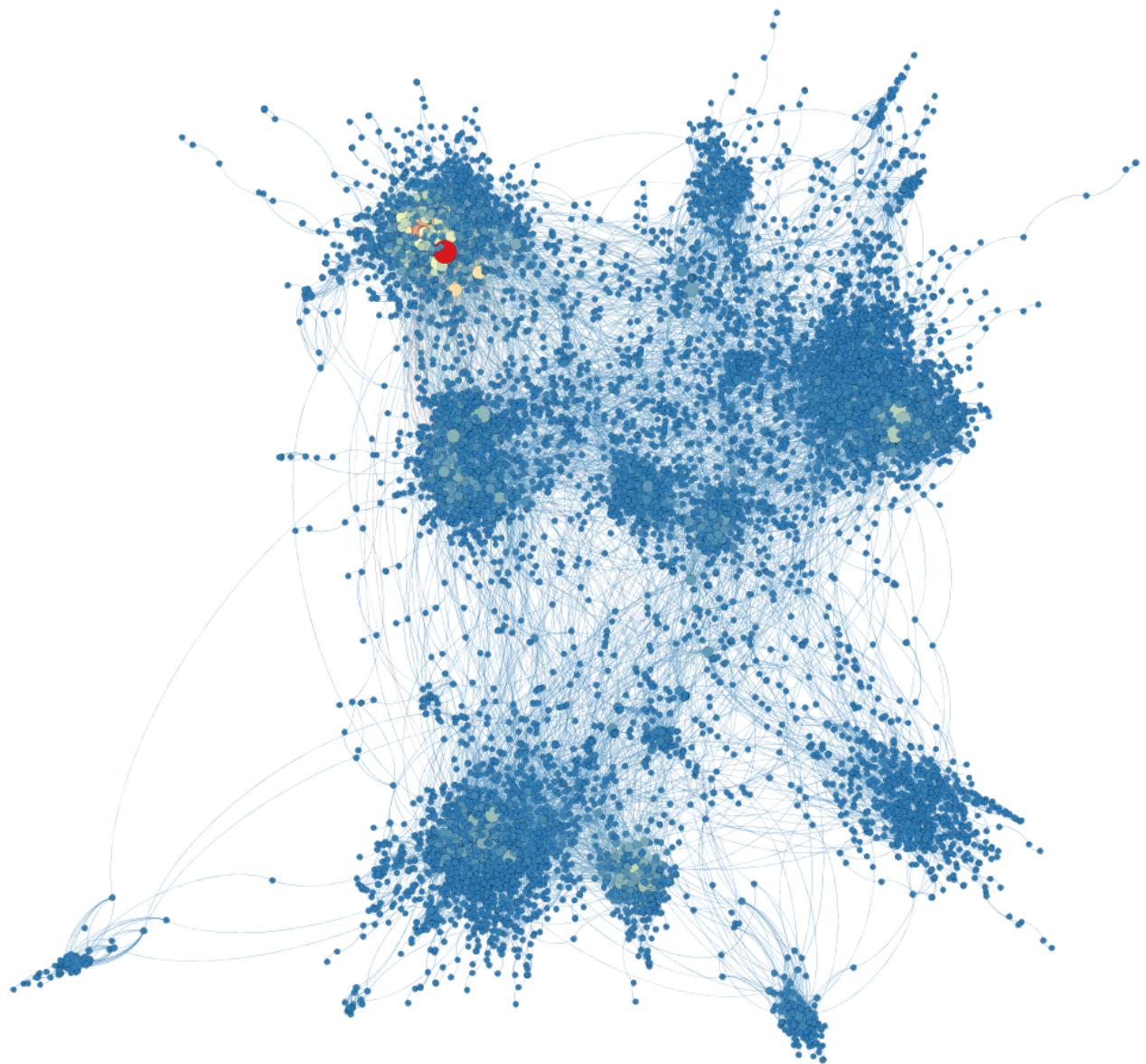


Figura 24: A mayor tamaño de nodo mayor grado. Azul implica menor vector propio, rojo más.

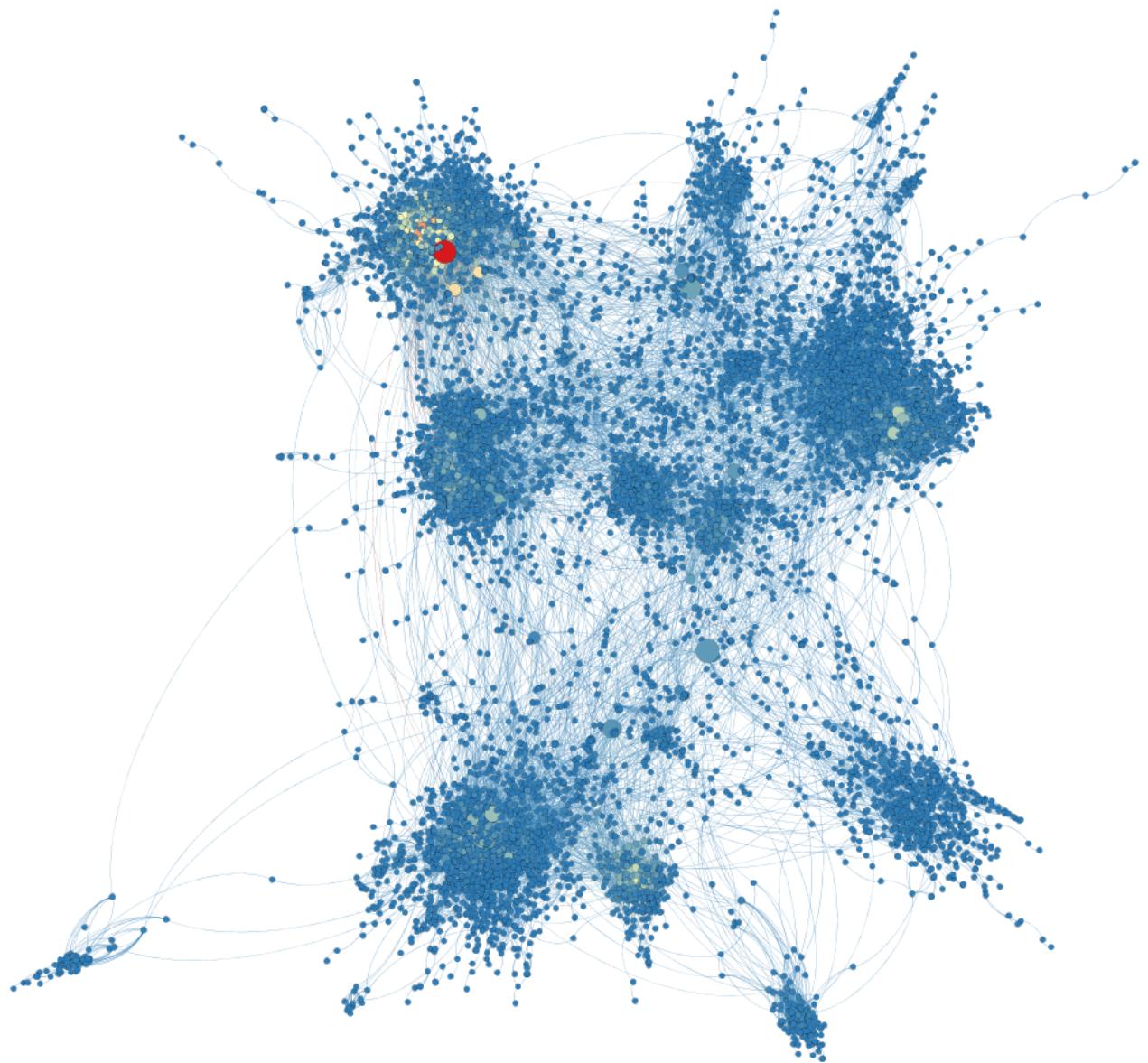


Figura 25: A mayor tamaño de nodo mayor intermediación. Azul implica menor vector propio, rojo más.