

# Práctica Minería de Medios Sociales.

Curso: 2018/2019. Néstor Rodríguez Vico. DNI: 75573052C - nrv23@correo.ugr.es

11 de mayo de 2019

## 1. Introducción.

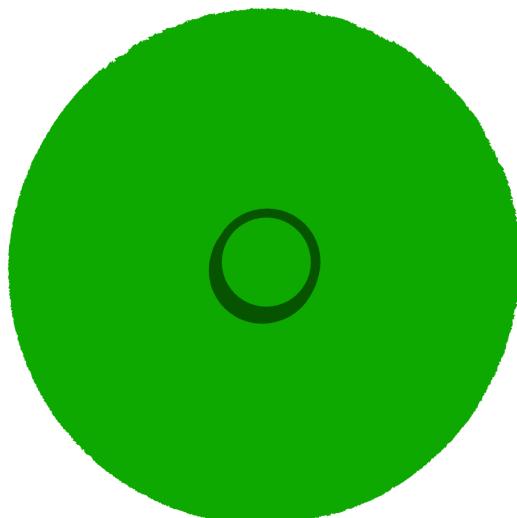
Con una tarea como la de predecir el tiempo de viaje de los taxis, puede ser interesante observar y estudiar la estructura de calles subyacente de la ciudad. El análisis de la red puede permitirnos entender por qué ciertos viajes en taxi toman más tiempo que otros, dadas algunas propiedades básicas de la red. El conjunto de datos (el cual se encuentra en el siguiente enlace: <https://www.kaggle.com/crailtap/street-network-of-new-york-in-graphml/>) usado contiene un gráfico para la red de calles para el área de Manhattan para realizar pruebas rápidas de su análisis. Cada nodo representa unas coordenadas en el mapa y hay una relación entre el nodo  $i$  y el nodo  $j$  si hay una ruta entre ambos nodos. Nos encontramos ante una red dirigida ya que, según las normas de circulación, poder ir del punto  $i$  al punto  $j$  no garantiza que se pueda ir del punto  $j$  al punto  $i$ . Dicha red ha sido obtenida con el software *osmnx* (<https://geoffboeing.com/publications/osmnx-complex-street-networks/>).

Aclaración: se ha usado Gephi para la realización de esta práctica pero, también se ha usado *networkx*, un módulo de Python, para obtener ciertos gráficos y métricas. Los scripts y el código relacionados con esto no se han entregado ya que no es el ámbito de la práctica pero, si fuese necesario, no dude en contactar conmigo y se los mando.

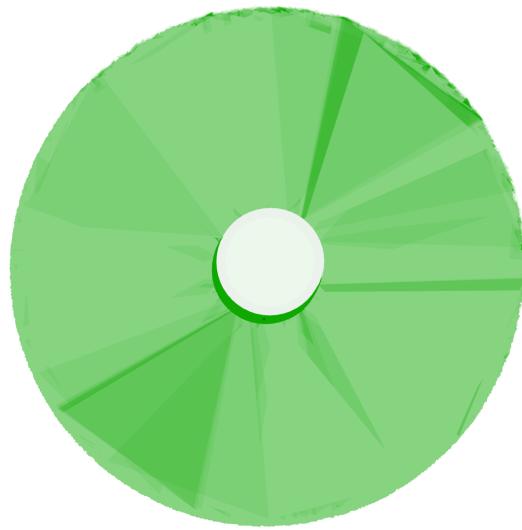
## 2. Análisis.

### 2.1. Análisis Básico de la Red.

Lo primero que vamos a hacer es mostrar la red:



Parece que no hay nada, pero si ponemos el ratón por encima, podemos ver que hay enlaces y nodos:



Para ver la red de una manera más correcta vamos a aplicar el algoritmo *Fruchterman Reingold* y el algoritmos *Force Atlas 2*:

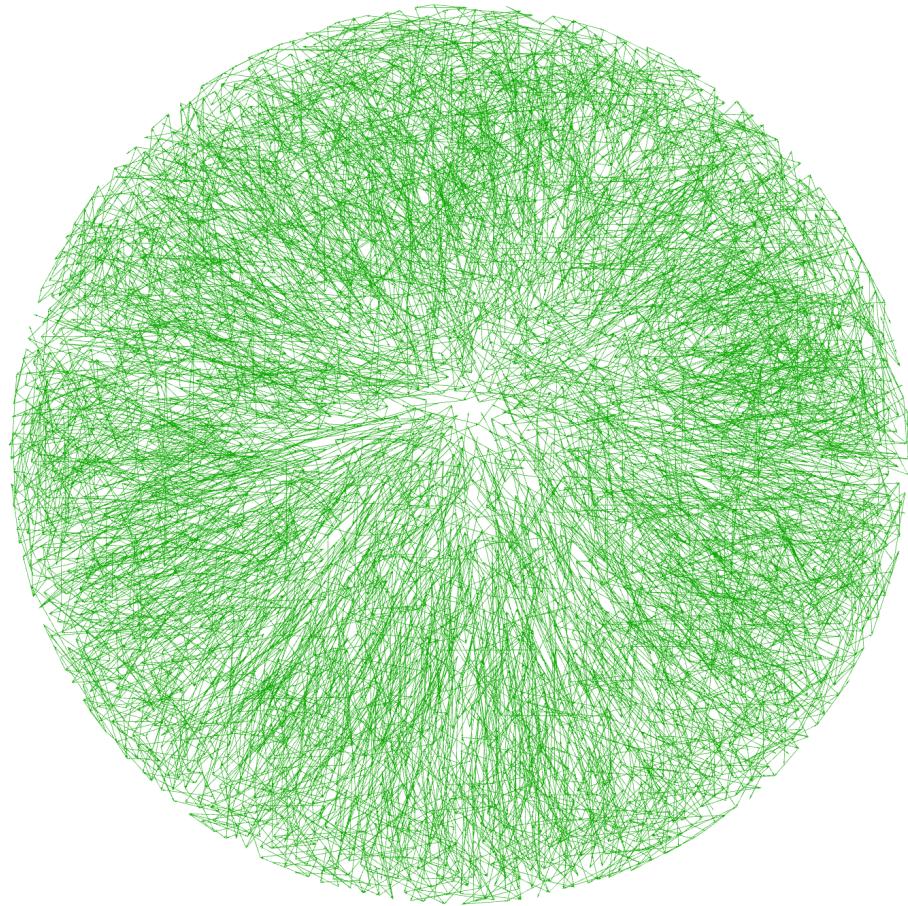


Figura 1: Visualización con *Fruchterman Reingold*.

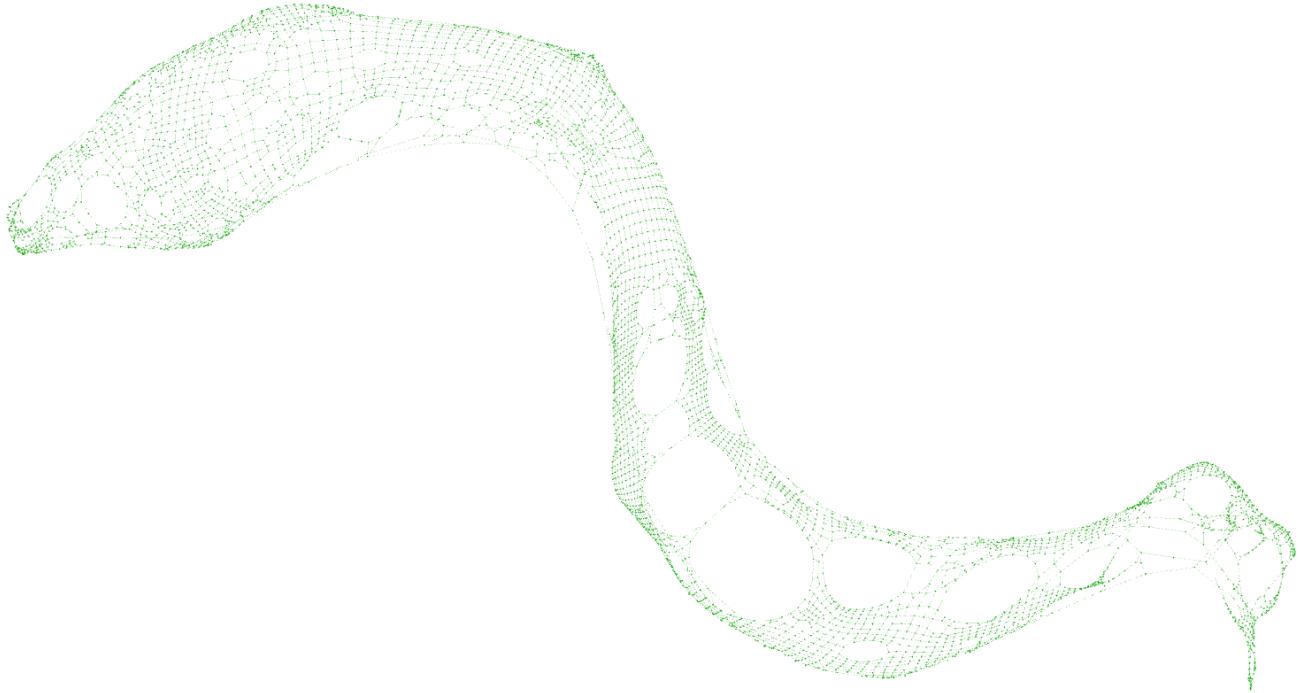


Figura 2: Visualización con *Force Atlas 2*.

A continuación, vamos a mostrar una tabla con ciertos valores generales de la red:

<b>Número de nodos: N</b>	4426
<b>Número de enlaces: L</b>	9626
<b>Número máximo de enlaces <math>L_{max} = N * (N - 1)</math>:</b>	19585050
<b>Densidad: D</b>	0.0004914
<b>Grado medio: <math>\langle k \rangle</math></b>	2.175
<b>Diametro: <math>d_{max}</math></b>	126
<b>Distancia media: <math>\langle d \rangle</math></b>	44.826846
<b>Distancia media aleatoria <math>\langle d_{aleatoria} \rangle</math></b>	1.09245
<b>Coef. de clustering medio <math>\langle C \rangle</math></b>	0.023
<b>Coef. de clustering medio aleatoria <math>\langle C_{aleatoria} \rangle</math></b>	0.491414
<b>Número de componentes conexas</b>	1
<b>Número nodos componente gigante</b>	4426
<b>Porcentaje con respecto a red total</b>	100 %
<b>Número enlaces componente gigante</b>	9626
<b>Porcentaje con respecto a red total</b>	100 %

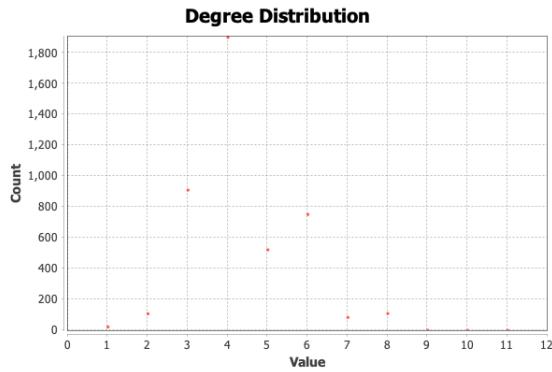
Para calcular la distancia media de una red aleatoria equivalente, he aplicado la siguiente fórmula:

$$\langle d_{aleatoria} \rangle = \frac{\log N}{\log \langle k \rangle} = \frac{\log 4426}{\log 2,175} = 1,09245$$

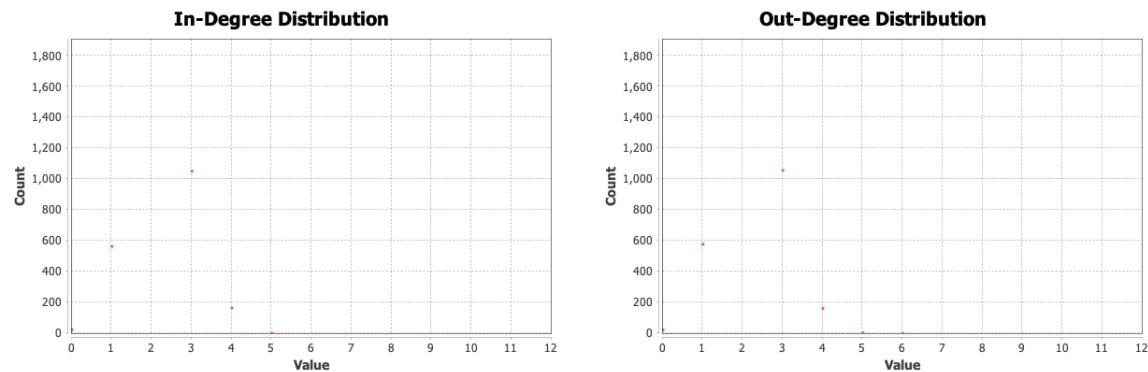
Para calcular el coeficiente de clustering medio de una red aleatoria equivalente, he aplicado la siguiente fórmula:

$$C = \frac{\langle k \rangle}{N} = \frac{2,175}{4426} = 0,491414$$

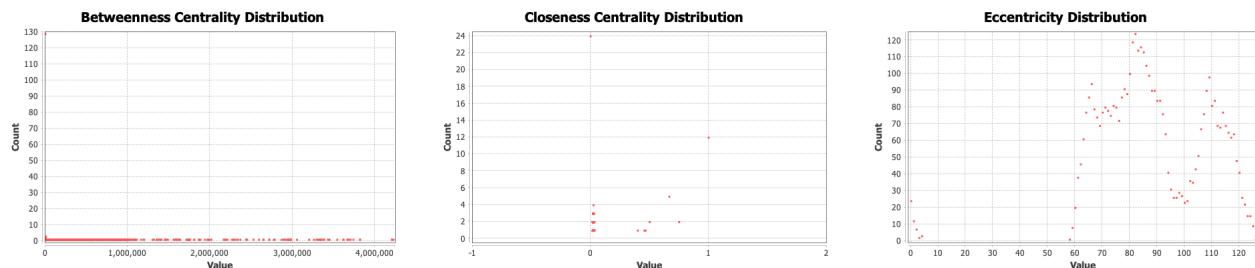
A continuación, vamos a ver la distribución de grados:



Dado que se trata una red dirigida, vamos a ver las distribuciones de grado tanto de entrada como de salida:



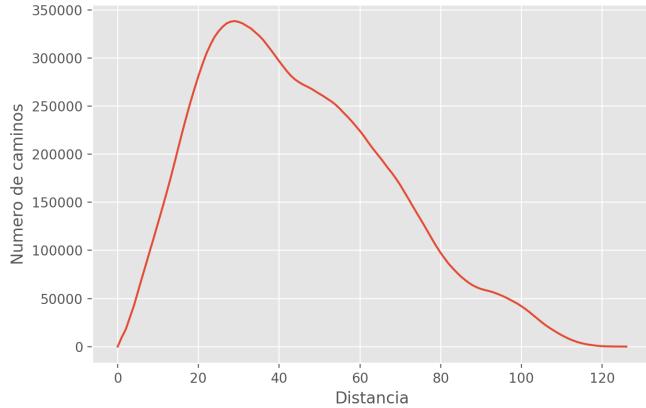
A continuación vamos a mostrar unas gráficas sobre tres medidas relacionadas con la distancia: *Betweenness* (intermediación), *Closeness* (cercanía) y *Eccentricity* (excentricidad).



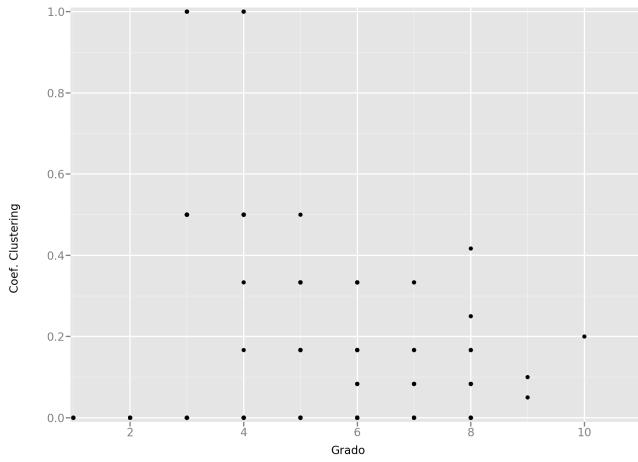
Otra gráfica de distancia interesante es la siguiente <sup>1</sup>:

---

<sup>1</sup>La idea de representar estas dos medidas juntas ha sido sacada del paper (figura 3) de *Renato Fabbri* nombrado en la bibliografía.



Finalmente, vamos a mostrar una gráfica sobre la distribución del grado de clustering:



Podemos concluir con un par de comentarios:

- Las distribuciones de grados de nuestra red nos indican que esta no sigue la llamada “Ley de la Potencia”. También podemos observar que no hay hubs claramente notables, ya que no hay una cantidad diferenciada de nodos con un grado más alto que los demás.
- Si observamos la última gráfica de distancias, podemos ver que el mayor número de caminos lo tenemos para una distancia en torno a 30. La gráfica no es simétrica del todo, por eso la distancia media es un poco más alta de 30.
- Si observamos la gráfica de clustering, podemos ver que no hay muchos nodos con un alto coeficiente de clustering y con un grado alto. Esto nos hace ver que las conexiones del área de Manhattan son bastante irregulares.
- Podemos ver que se trata de una red de mundo pequeño observando la gráfica de distancia. En dicha gráfica se representa el número de caminos de una distancia dada. Podemos ver que la mayoría de los caminos tienen una distancia pequeña, cercana a la distancia media. También vemos que el número de caminos que tiene una distancia grande son mínimos. Por lo tanto, podemos decir que estamos ante una red de mundo pequeño.

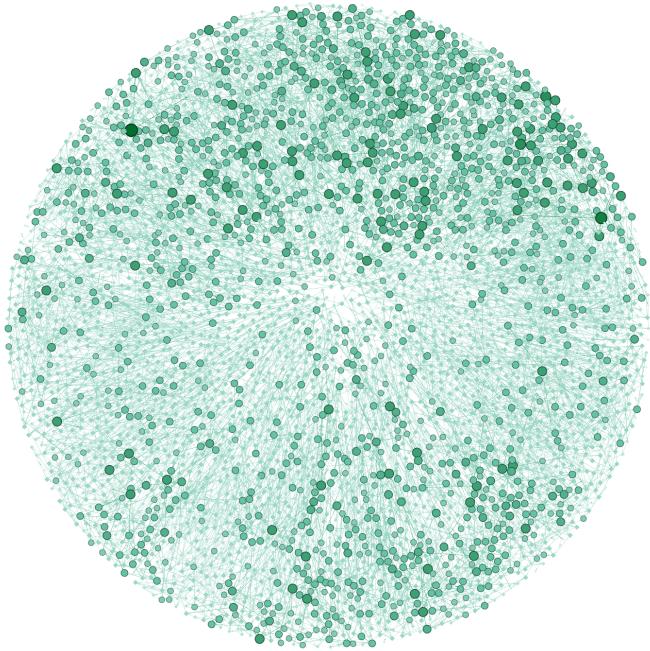
## 2.2. Estudio de la Centralidad de los Actores.

Los cinco actores mas relevantes para cada medida son los siguientes:

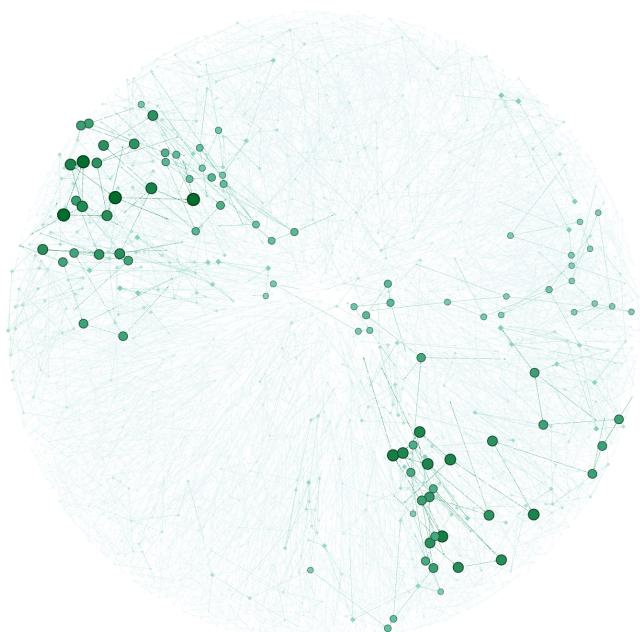
Centralidad de Grado	Centralidad de Intermediación	Centralidad de Cercanía	Centralidad de Vector propio
2512378850: 11	42431099: 0.2168	486819531: 1	42427867: 1.0
42431257: 10	587812578: 0.2154	42423847: 1	42427870: 0.976
42429215: 9	42441382: 0.2148	371225088: 1	42437923: 0.964
42429226: 9	42441310: 0.2145	588455736: 1	42431242: 0.949
42422283: 8	42436719: 0.1948	3785586748: 1	42431252: 0.914

- Debemos tener cuidado con el grado, ya que como hemos visto en clase de teoría, se trata de una medida bastante local.
- La intermediación capta la correduría de la información por la estructura de la red. Obtiene un mayor valor para los nodos por los que pasen más caminos mínimos por él. Estos nodos son los que se encargan de hacer de puente entre comunidades y, por lo tanto, son usados para detectar comunidades.
- Como hemos visto en las transparencias de teoría, la cercanía es una forma de medir la centralidad, la cual plantea que el hecho de que puede no ser tan importante tener muchos amigos directos ni estar situado “entre” otros actores. En este caso, se le da importancia a “estar en medio de las cosas”, no demasiado lejos del centro, para lo cual no es necesario estar en una posición de correduría. La suma de las distancias geodésicas (distancias de los caminos mínimos) para cada actor es la lejanía de dicho actor al resto. La inversa de dicha suma es la medida de cercanía.
- La centralidad de vector propio tiene como base la idea de que la centralidad de un nodo depende de cómo de centrales sean sus nodos vecinos. La idea es que el poder y el status de un actor (ego) se define recursivamente a partir del poder y el status de sus vecinos (alters). Es una versión más elaborada de la centralidad de grado al asumir que no todas las conexiones tienen la misma importancia. Es como dice el dicho, “más vale calidad que cantidad”.

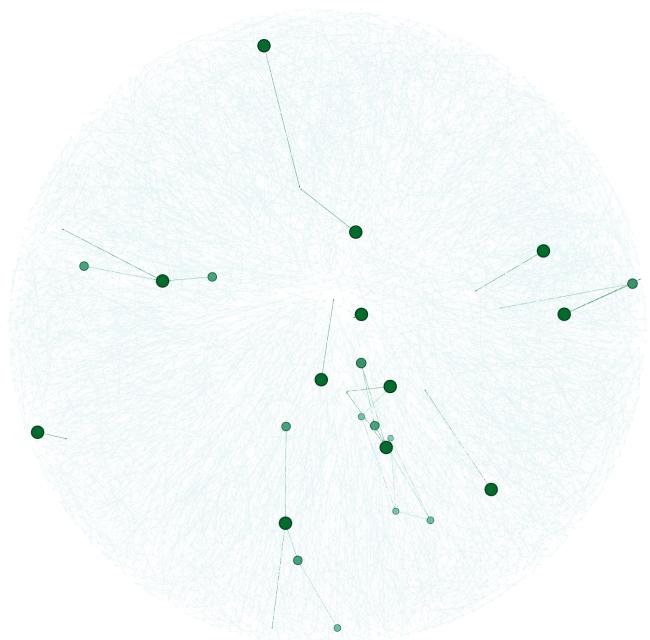
A continuación voy a mostrar la red destacando las medidas estudiadas en esta sección. Para ello, he aplicado el algoritmo *Fruchterman Reingold* para poder visualizar la red de una forma más práctica. A continuación, se ha cambiado el tamaño y el color de los nodos para representar las distintas métricas: a mayor tamaño y más oscuro es el color verde, mayor es la medida representada. Quiero remarcar un aspecto interesante, la medida de grado está bastante repartida, es decir, no hay nodos que destaque. Podemos ver como en las otras tres representaciones si hay nodos más destacados, sobre todo en la medida de centralidad de vector propio.



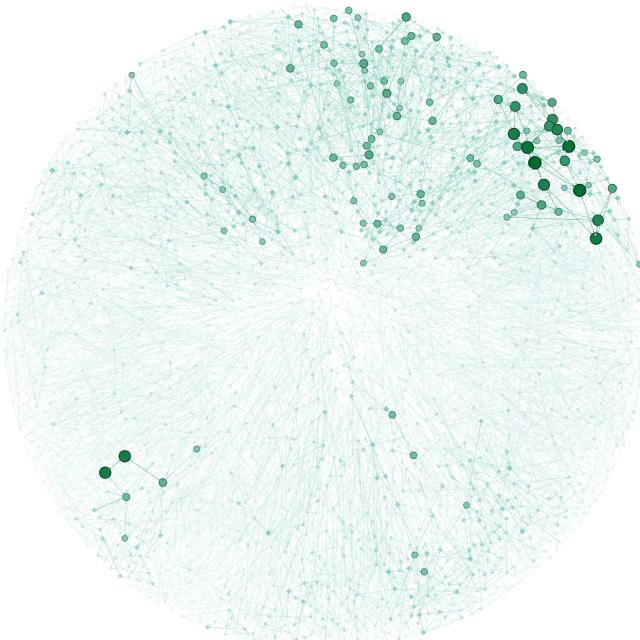
(a) Centralidad de grado.



(b) Centralidad de intermediación.

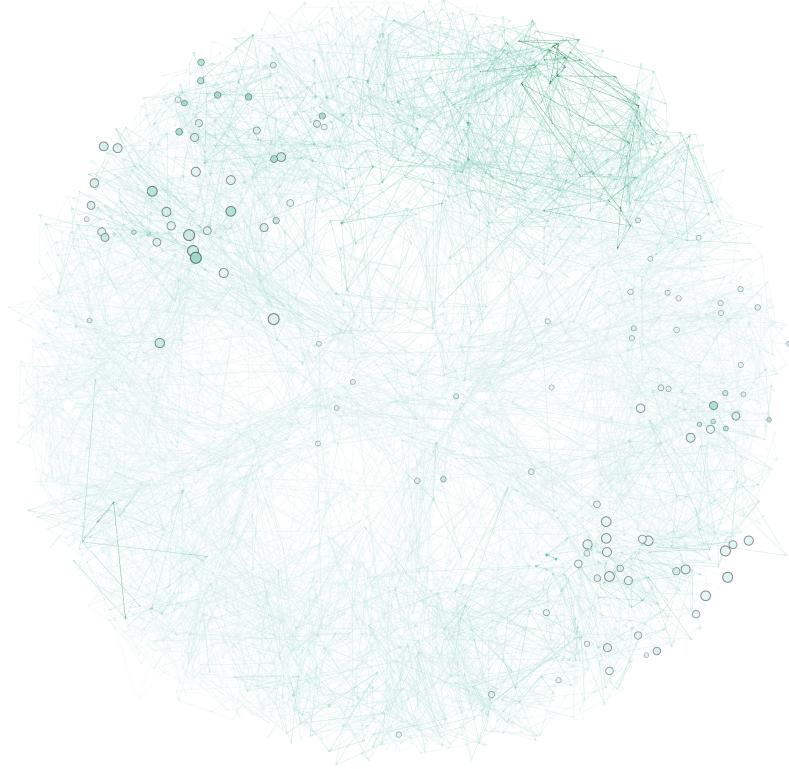


(c) Centralidad de cercanía.



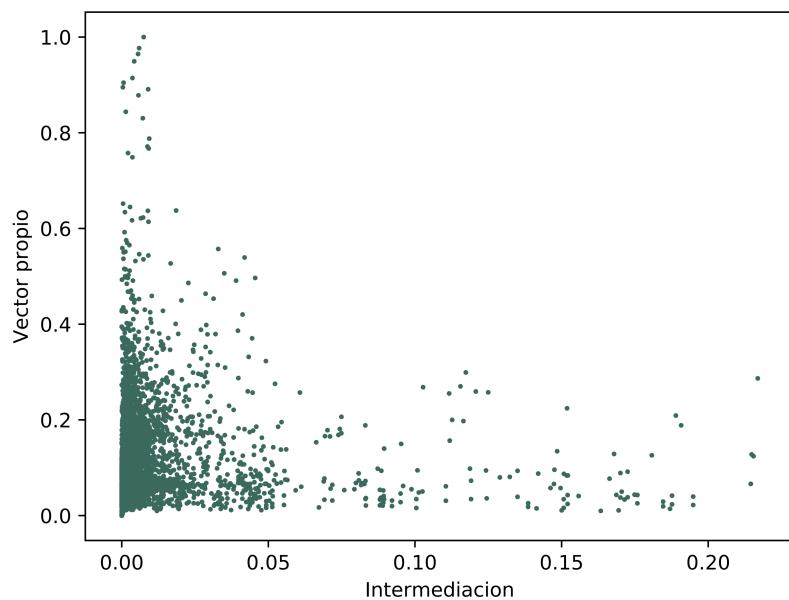
(d) Centralidad de vector propio.

Finalmente, tal y como se indica en el guión de prácticas, vamos a mostrar una gráfica en la que se visualizan dos de las medidas anteriores, la intermediación en el tamaño de los nodos y la centralidad de vector propio en el color de los mismos:



Cómo podemos ver, hay nodos en los que ambas medidas coinciden, como puede ser la zona de la izquierda donde se ven los nodos más grandes. Sin embargo, en la zona superior derecha, podemos ver que los nodos tienen un alto valor de la centralidad de vector propio (nodos muy oscuros) pero una intermediación bastante pequeña (nodos muy pequeños).

Finalmente, vamos a mostrar un gráfico que representa los valores de dos de las medidas para todos los actores de la red en ejes de coordenadas como los estudiados en la Sesión I.2. En mi caso, voy a representar las mismas medidas que he mostrado en el gráfico anterior:

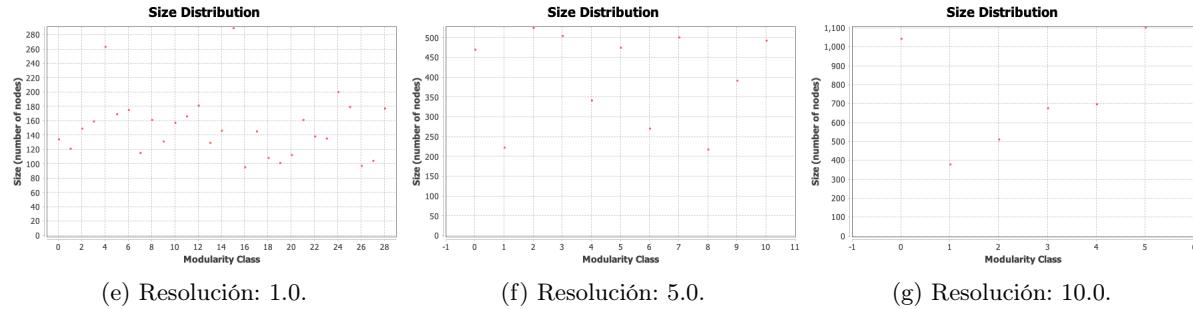


Como podemos ver, la mayoría de los nodos tienen un valor de vector propio muy bajo y un valor de

intermediación muy bajo. También podemos ver que los nodos que tienen un valor alto de vector propio tiene un valor bajo de intermediación y viceversa.

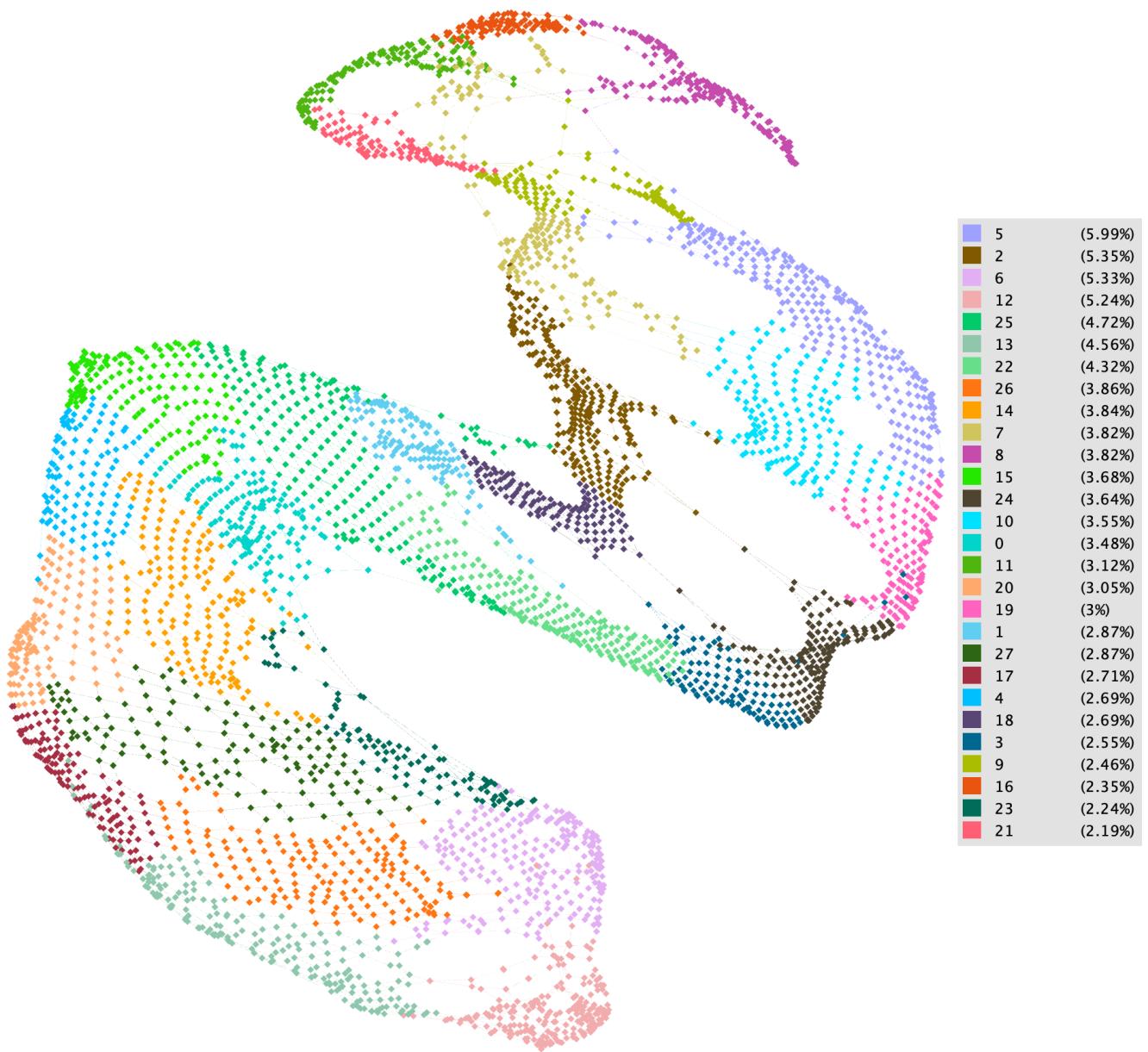
### 2.3. Detección de Comunidades.

Para la detección de comunidades se ha aplicado el método de *Lovain* con distintos valores para el parámetro resolución. Los resultados obtenidos son los siguientes:

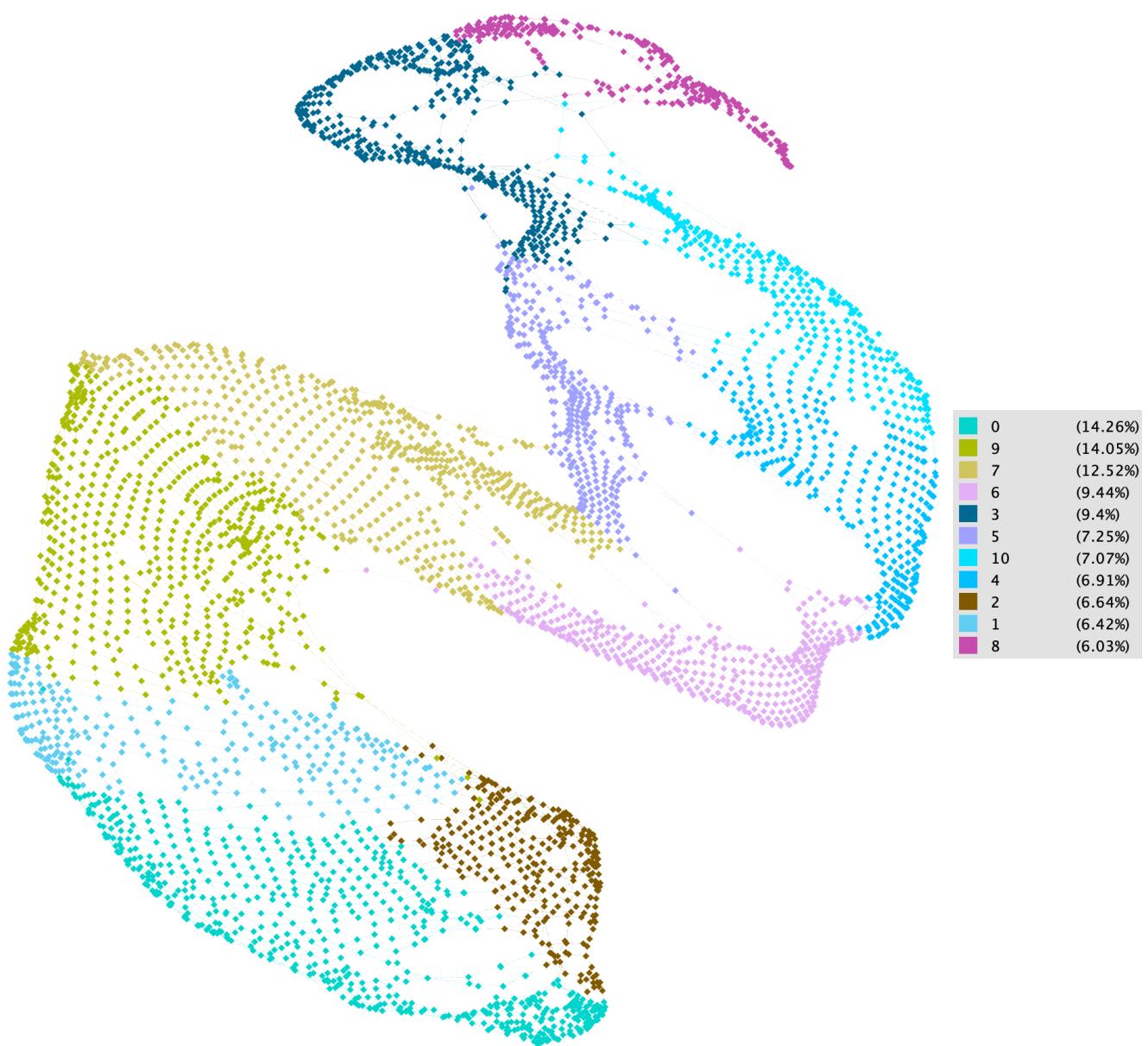


Resolución	Número de comunidades	Modularidad	Modularidad con resolución
1.0	29	0.906	0.906
5.0	11	0.874	4.764
10.0	6	0.797	9.662

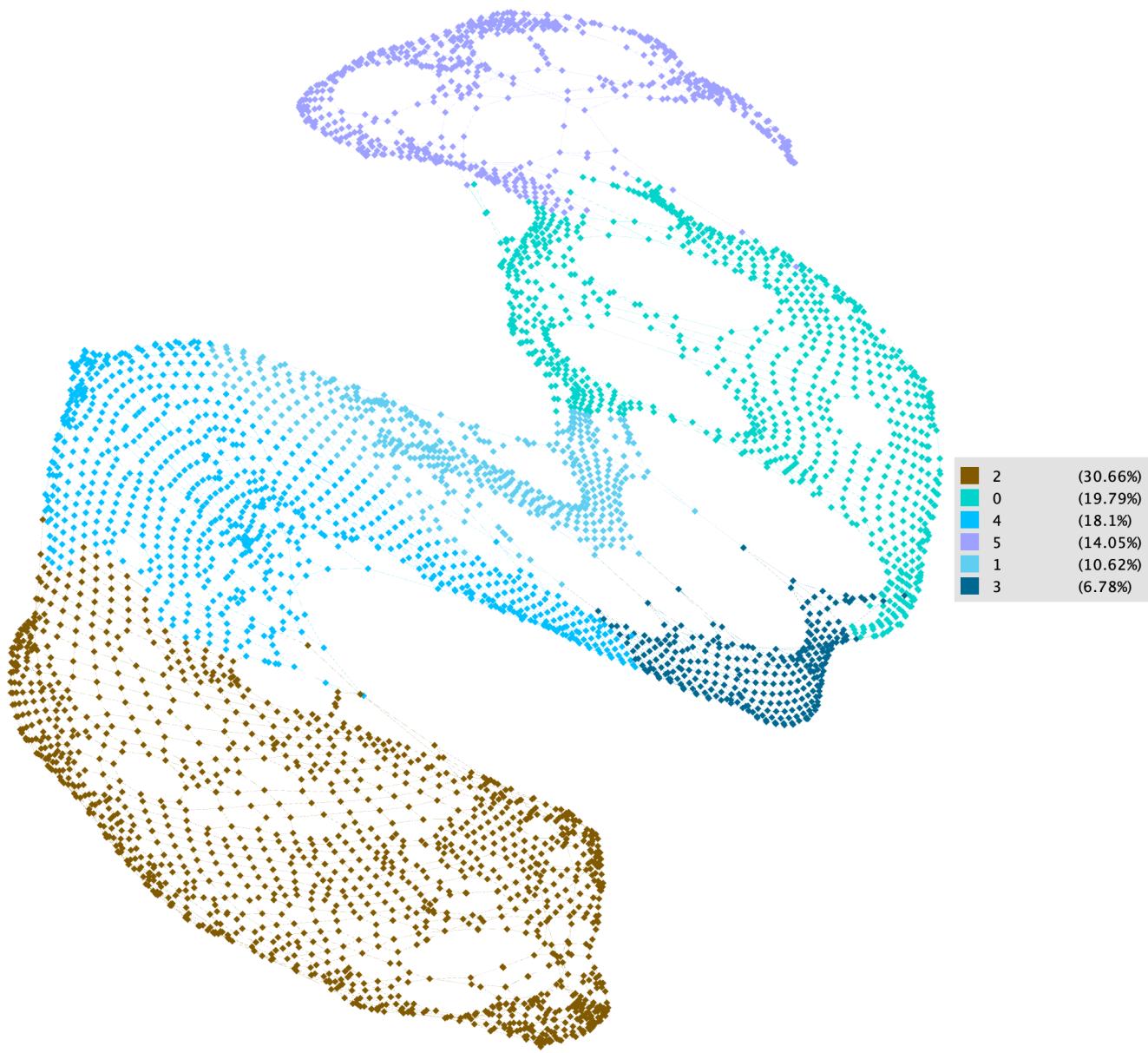
Cómo podemos ver, aumentar el valor de resolución supone reducir el número de comunidad y, por lo tanto, reducir el valor de la modularidad. A su vez, aumenta la modularidad con resolución de la red. A continuación, voy a mostrar unas imágenes que representan la red en función de la comunidad a la que pertenecen. Primero vamos a ver los resultados obtenidos para una resolución de *1.0*:



A continuación, vamos a ver los resultados obtenidos para una resolución de *5.0*:



Finalmente, vamos a ver los resultados obtenidos para una resolución de  $10.0$ :



Para la visualización de las comunidades he usado el algoritmo *Force Atlas 2*, ya que permite ver con mucha claridad las comunidades detectadas en la red. Como podemos ver, no hay comunidades claras como hemos estudiado en la asignatura, ya que ninguna frontera de las mismas es totalmente clara. Sin embargo, podemos ver como las comunidades representan zonas contiguas de Manhattan. En el problema concreto que estamos estudiando, esta idea de comunidades se podría utilizar para definir zonas por las cuales se pueden mover los taxis y poner un precio al servicio según la zona en la que se mueven o según la zona de origen y destino.

#### 2.4. Gráficos adicionales.

Los gráficos adicionales se han mostrado a lo largo de la memoria en su sección correspondiente.

### 3. Bibliografía.

- networkx: Herramienta usada para realizar ciertas gráficas.

- graph-tool: Herramienta usada para realizar ciertas gráficas.
- On the evolution of interaction networks: primitive typology of vertex and prominence of measures  
- Renato Fabbri, Vilson Vieira da Silva Junior, Ricardo Fabbri y Osvaldo N Oliveira
- Gephi.