I Curos Internacional Virtual “Cátedra de la Provincia UHU”

Social Media e Internet: Extracción de Datos Masivos (Big Data), Procesado, Análisis y Visualización con ayuda de R, Python y Gephi.

Evaluación Módulo 5: Análisis de redes y visualización

Realizado por:
Álvaro Esteban Muñoz

Índice

Introducción

En las próximas páginas se detalla el estudio de una red basada en datos extraídos de twitter mediante la librería "tweepy", así como su visualización en **Gephi.**

A lo largo de este informe se comenta más detalladamente como ha sido realizada la obtención de los datos. También se analiza la red en profundidad y se visualizan dos grafos diferentes destacando indicadores diferentes.

Datos generales de la red elegida

La red elegida se ha montado a partir de una serie de datos extraída de Twitter mediante la librería "tweepy" para **Python**. El script usado para extraer dichos datos ha sido realizado por mi mismo y viene en la misma carpeta que este informe. Los parámetros usados para extraer los datos son los siguientes:

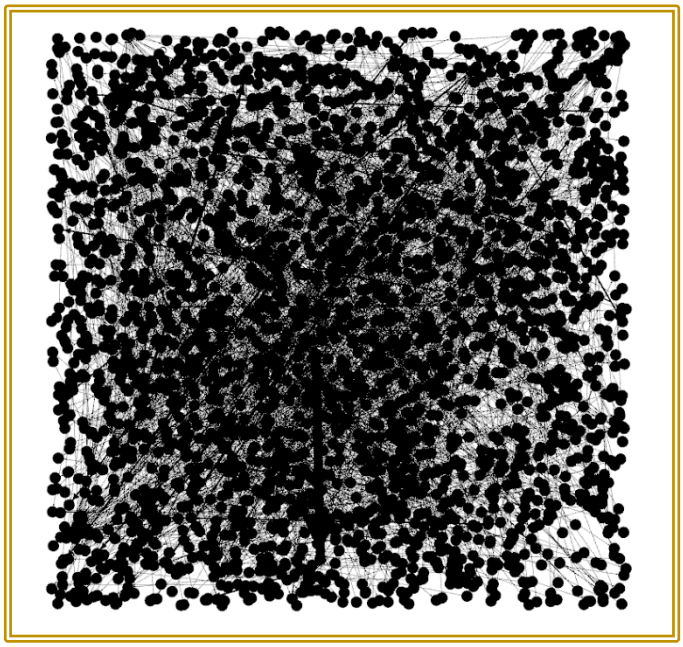
- *q = "covid" or "coronavirus"*, donde **q** denota el valor de búsqueda por el que se seleccionarán tweets.
- *lang = None*, este parámetro no se ha indicado para que no se seleccione dependiendo del lenguaje.

La red está compuesta de 2972 tweets (Se esperaba conseguir 3000). El grafo formado es dirigido y las aristas indican los retweets (saliente = retweet, entrante = retweeted). Está formado por 2807 nodos y 1868 aristas.

En el apartado a continuación se podrá observar el comportamiento de la red al destacar indicadores y aspectos diferentes de la misma.

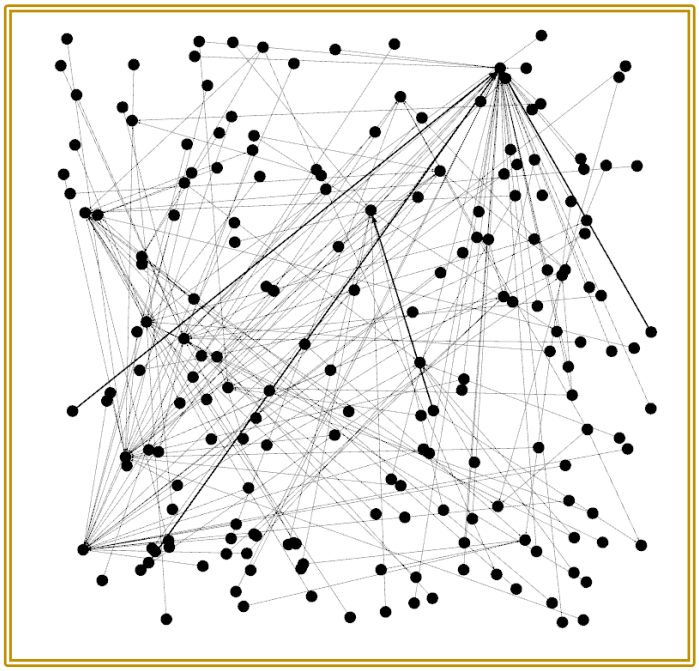
Análisis de la red de datos

Lo primero que vemos al cargar la red es lo siguiente:



Claramente se hace muy difícil sacar algo en claro de esto, por eso vamos a usar las herramientas que ofrece **Gephi** para transformarlo en algo más legible y de donde podamos extraer información útil.

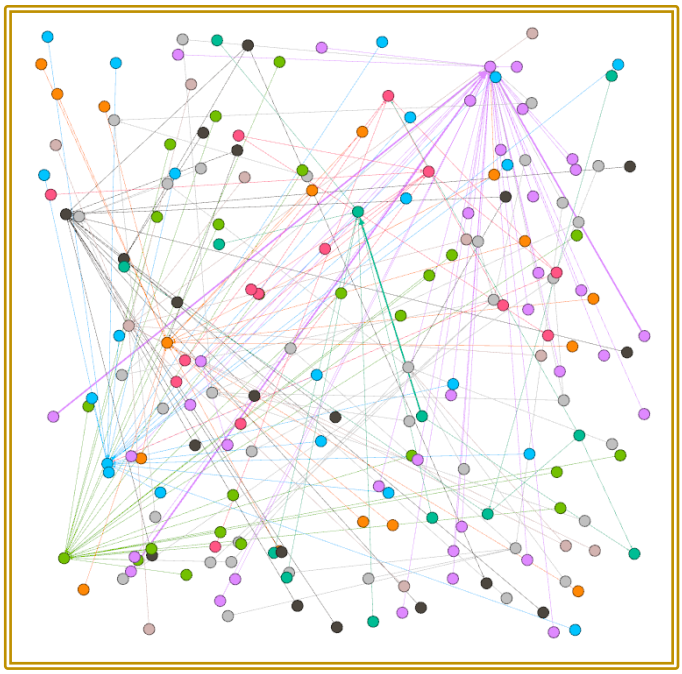
Para que la información que representa el grafo sea más legible y el análisis sea más consistente vamos a aplicar el filtro de **componente gigante**, esto eliminará los nodos aislados o inaccesibles y nos evitará el sesgo provocado por los mismos en los resultados.



Tras aplicar el filtro de componente gigante obtenemos un grafo mucho más fácil de manejar. Vamos a destacar dos indicadores:

- Grado medio ponderado
- Centralidad por intermediación

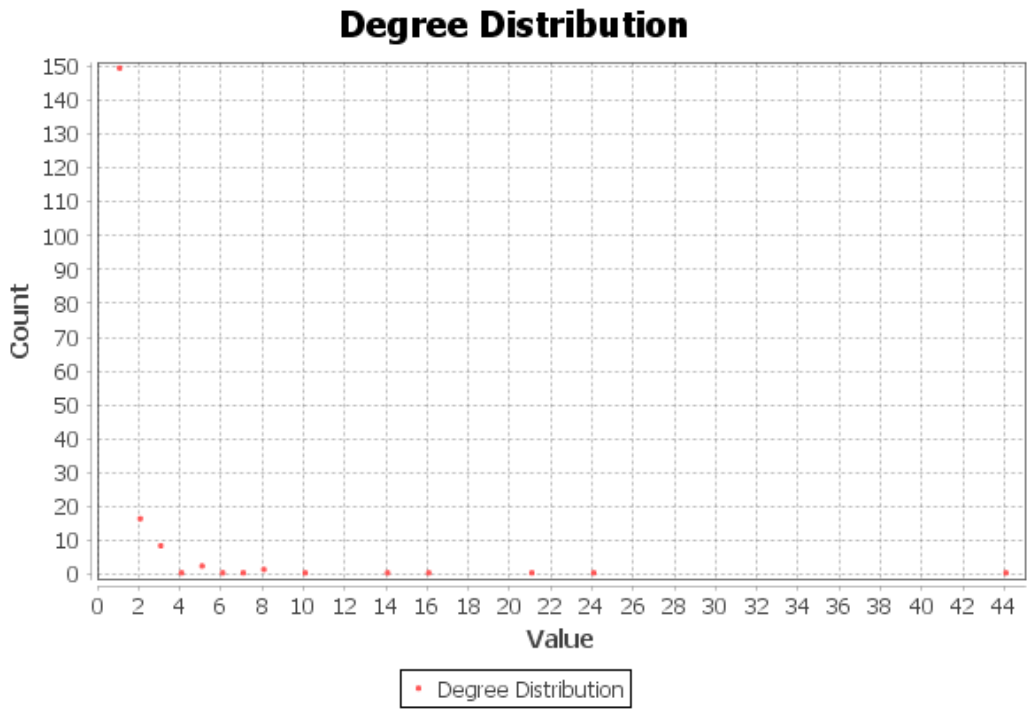
Con estos dos indicadores podremos saber la importancia de los nodos para el grafo, pudiendo reconocer los usuarios con más influencia y popularidad de la red, así como aquellos usuarios por los que fluye más la información, es decir, aquellos por los que suele pasar más información. También colorearemos los nodos por la modularidad para reconocer comunidades.

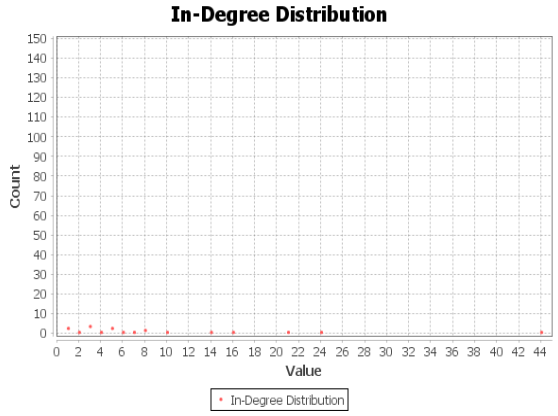


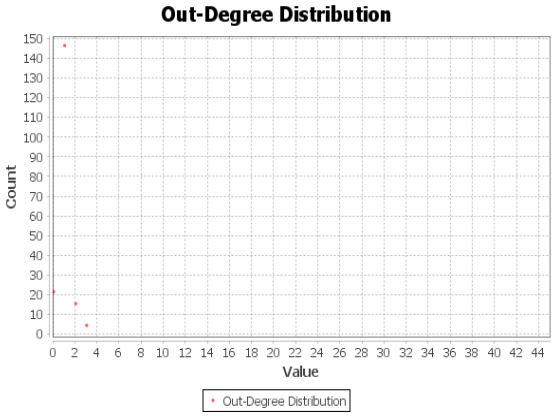
Grado medio ponderado

Gephi genera un informe con los resultados de calcular el indicador seleccionado junto a unas gráficas del mismo.

Average Weighted Degree: 1,021

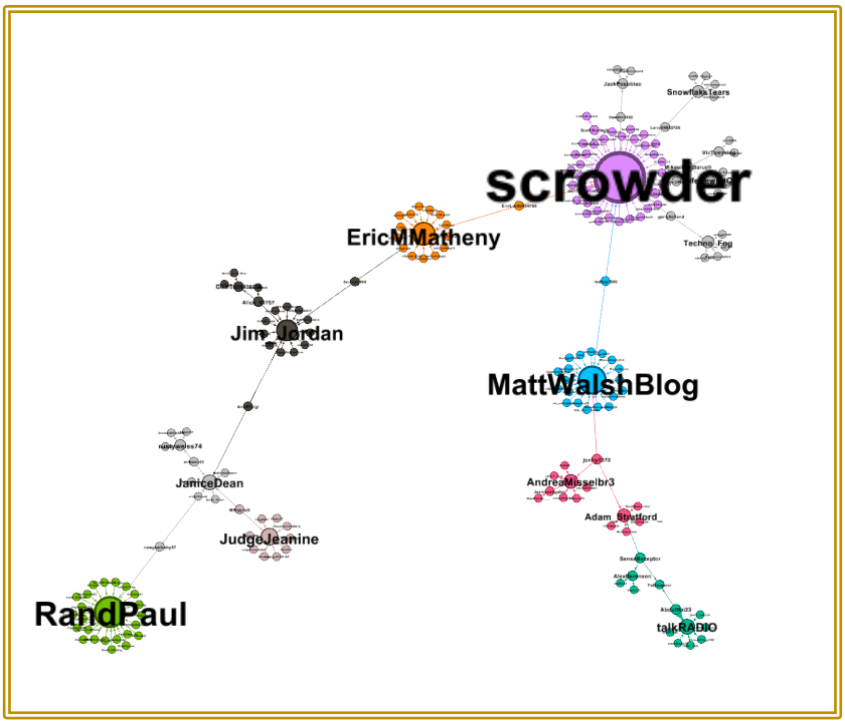






A la hora de visualizar el grafo destacaremos el grado medio ponderado en el tamaño de los nodos, aquellos nodos con más popularidad en influencia serán más grandes.

Usaremos el algoritmo de distribución "*Force Atlas 2*" para que nos agrume los nodos y sea más fácil de visualizar el grafo.



Podemos observar en el grafo una serie de comunidades formadas alrededor de unos cuantos usuarios concretos de los cuales vamos a investigar un poco más.

Podemos observar los nodos más importantes de la red en laboratorio de datos:

Id	Label	Inter...	links	links_in	links_out
scrowder	scrowder		44	44	0
RandPaul	RandPaul		24	24	0
MattWalshBl...	MattWalshBlog		21	21	0
EricMMatheny	EricMMatheny		16	16	0
Jim_Jordan	Jim_Jordan		14	14	0
JudgeJeanine	JudgeJeanine		10	10	0
JaniceDean	JaniceDean		8	8	0
talkRADIO	talkRADIO		8	8	0

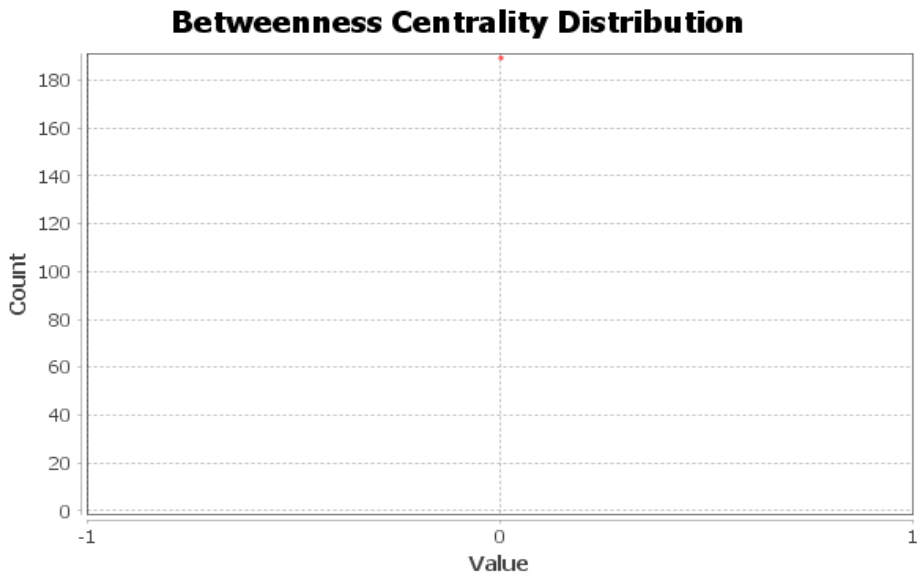
Podríamos ir a Twitter a buscar más información sobre alguno de los usuarios, podemos notar por qué **scrowder** maneja tanto flujo de información.

Vemos que es un creador de contenido que maneja un programa en directo, si quisiéramos hacer un análisis más detallado de este usuario podríamos averiguar su posicionamiento político y de ahí deducir que toda la información que maneja su comunidad será de una ideología similar puesto que los usuarios de Twitter tienden a retweetear aquello con lo que están de acuerdo.

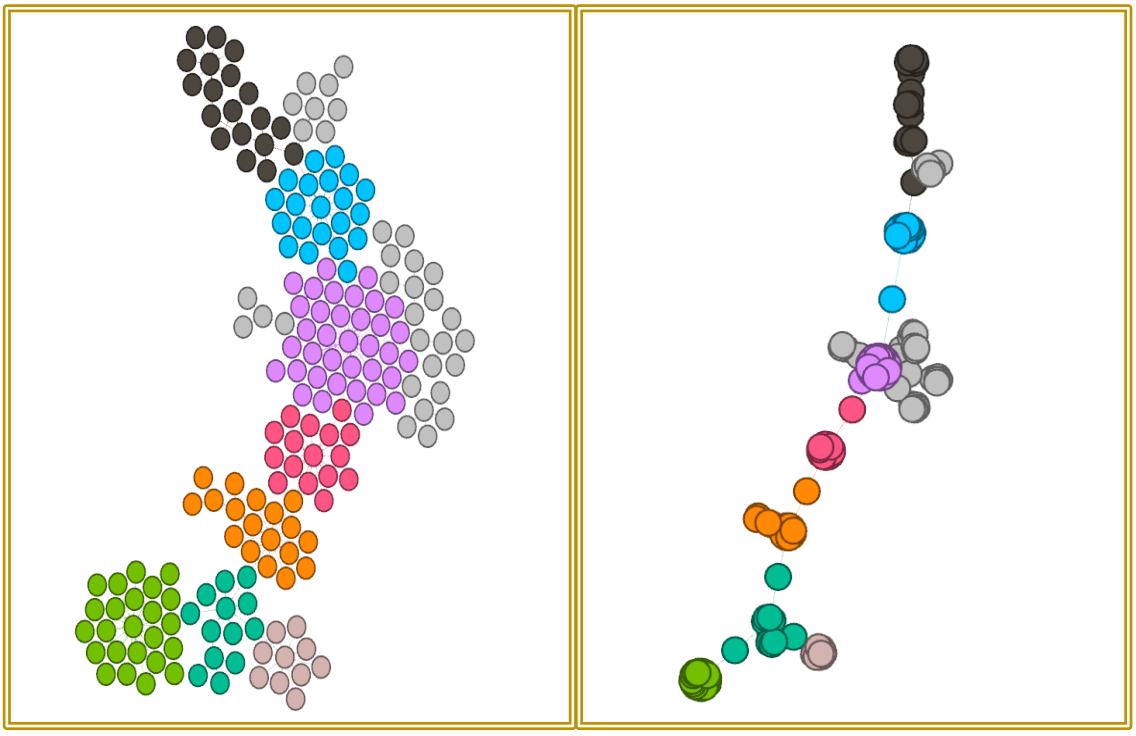
Centralidad por intermediación

El informe generado por Gephi es el siguiente:

Diameter: 1
Radius: 0Average Path length: 1.0



Si destacamos este indicador en la visualización del grafo obtendremos algo parecido a lo siguiente:



Aquí podemos observar que todos los nodos propagan la información con la misma influencia.

Conclusiones

En la red de datos extraídos se denotas los usuarios con más influencia respecto al coronavirus en los últimos 7 días. Debemos tener en cuenta que el análisis se ha realizado sobre una red de 3000 tweets sin concretar ninguna región en particular, es decir, los resultados pueden ser muy dispares y para hacer un análisis más consistente tendríamos que extraer muchos más datos y concretar algo más la población elegida y obtener así una muestra más pura.

No obstante, estos resultados son muy satisfactorios si los consideramos orientativos. Realizar un análisis sobre una masa de datos mucho más extensa nos aportaría unos resultados de mayor calidad, sin embargo, la metodología usada no variaría mucho de lo que ya se ha expuesto en este informe.