Análisis de Redes Sociales

Guillermo Jiménez Díaz (gjimenez@ucm.es)
Alberto Díaz (albertodiaz@fdi.ucm.es)

Curso 2018 - 2019

Práctica 1: Análisis de una red con Gephi

La práctica tiene por objetivo la creación y el posterior análisis de una red. Para la realización de la práctica se proponen dos posibles conjuntos de datos:

- La creación de la red de usuarios de Twitter más importantes en distintos países.
- La creación de una red de interacción a partir de la información de los personajes que aparecen en las escenas de la serie Los Simpsons.

Independientemente del conjunto de datos utilizado, la realización de la práctica se dividirá en dos partes bien diferenciadas:

1. Generación de la red: Los conjuntos de datos proporcionados **no están en forma de grafo** por lo que será necesario procesarlos para que sean visualizables en Gephi.
2. Visualización y extracción de información: Los datos procesados serán cargados en Gephi para poder visualizar la red y extraer información adicional sobre los nodos y la estructura de la red en sí misma.

1. Generación de la red

1.1. Twitter

En primer lugar cada estudiante deberá analizar las redes de usuarios de Twitter de varios países. Se deberán discutir las diferencias y similitudes entre redes de al menos dos países distintos.

Los archivos de información fueron generados por los miembros del proyecto Twitter Alliance, haciendo uso de la información contenida en https://twittercounter.com y el acceso a la API de Twitter. Para cada país se proporcionan dos archivos:

- `Top100_<pais>_friendships_users.txt`: Archivo en el que se almacena la información de los 100 usuarios de Twitter más importantes del país junto con un conjunto de estadísticas adicionales:

- `country`: País del usuario.
- `following`: Número de usuarios a los que sigue.
- `followers`: Número de usuarios que le siguen.
- `tweets`: Número de tweets publicados.
- Otra información sobre el lugar, latitud y longitud especificados en el perfil, número de favoritos y de listas de las que dispone.

Cada línea del archivo tiene el siguiente aspecto (el texto entre `[]` significa que puede estar o no disponible:

```
@username: country following followers tweets [otra información]
```

- `Top100_<pais>_friendships.txt`: Es una lista con los seguidores (de la lista del Top 100) que siguen a ese usuario. El número de seguidores no se conoce a priori. Cada línea del archivo tiene el siguiente aspecto:

```
@username: @username1 @username2 ... @usernameN
```

Lo siguiente a realizar es crear un programa (en el lenguaje que se prefiera) para procesar los datos de estos archivos de texto y crear la lista de nodos y de aristas en un archivo cuyo formato sea uno de los formatos que Gephi es capaz de cargar. Los nodos representan cada uno de los usuarios de Twitter del Top 100 mientras que cada arista representa que un usuario sigue a otro. Hay que tener en cuenta que este grafo es **dirigido** y que todas las aristas tienen un peso de 1.

Por ejemplo, si creamos un fichero de nodos y un fichero de aristas en formato CSV compatible con Gephi, el aspecto de ambos archivos será algo parecido a lo siguiente:

Lista de nodos:

```
Id,Label
0,@Carles5puyol
1,@SergioRamos
2,@RafaelNadal
3,@Rubiu5
```

Lista de aristas:

```
Source,Target,Type
0,2,Directed
1,2,Directed
0,3,Directed
1,3,Directed
```

2. Análisis y visualización de las redes generadas

La segunda parte de la práctica consiste en la carga de los grafos *de al menos dos países* en Gephi, para su posterior análisis y visualización. Es recomendable que juguéis con los filtros y los algoritmos de visualización de Gephi para conseguir una información que seáis capaces de interpretar adecuadamente. Podéis generar múltiples redes en función de los filtros utilizados para sacar diferentes conclusiones.

Para cada una de las redes generadas hay que generar un informe con las siguientes secciones.

2.1. Medidas globales de la red

El informe de cada red ha de incluir las siguientes medidas[1]:

- Número de nodos N
- Número de enlaces L
- Densidad de la red.
- Grado medio ($\langle k \rangle$)
- Diámetro (d_{max}) y distancia media ($\langle d \rangle$).
- Coeficiente medio de clustering ($\langle C \rangle$)

También hay que incluir los histogramas con la distribución de grados. Estos histogramas se generan al calcular las estadísticas correspondientes.

2.2. Medidas de centralidad

En este caso se incluirán las siguientes métricas:

- Intermediación
- Cercanía
- Excentricidad
- PageRank (hay que indicar los parámetros usados para su cálculo)
- Centralidad de vector propio (hay que indicar los parámetros usados para su cálculo)

2.3. Conectividad

Con respecto a la conectividad de la red, el informe ha de especificar la siguiente información:

- Número de componentes conexas.
- Modularidad y número de comunidades. Es recomendable jugar con los parámetros del cálculo de la modularidad para intentar encontrar comunidades que tengan sentido en el contexto de la red que estamos analizando. Así mismo, hay que indicar qué parámetros se han usado para calcular la modularidad que se ha considerado representativa.

[1] Las métricas pedidas están accesibles o hay que calcularlas usando las opciones correspondientes en la ventana `Estadísticas`.

2.4. Visualización

Hay que incluir al menos una imagen, generada en Gephi, que apoye las principales conclusiones obtenidas de las redes. Por ejemplo, se pueden utilizar distintos colores para los nodos que están en distintos módulos (comunidades) y distintos tamaños para reflejar una determinada medida de centralidad. Hay que buscar la configuración que más información pueda dar de la red analizada.

Los aspectos estéticos de la visualización se dejan al parecer del propio estudiante, que debe probar las distintas variantes de algoritmos de layout implementados en Gephi y de parámetros para determinar cuál le proporciona la visualización que resulte más informativa. En todo caso, **es fundamental incluir en el informe el algoritmo utilizado y los parámetros empleados para cada una de las imágenes que se incluyan**. Puede haber varias imágenes con layout distintos si se considera relevante.

2.5. Opciones de filtrado

Después de hacer un análisis global de la red es recomendable analizar algunas partes de la red de manera más detallada (por ejemplo, nodos con mayor grado, comunidades concretas, etc.). Es imprescindible indicar cuáles son los filtros aplicados a la red y cuáles han sido los parámetros de dichos filtros. También es recomendable indicar con qué proporción de nodos y aristas nos hemos quedado tras la aplicación del filtro (información que aparece en Gephi).

3. Discusión

Esta sección consiste en realizar un análisis de los datos obtenidos, extrayendo información relevante de los datos. No consiste en enumerar de nuevo los datos sino en extraer conclusiones. Por ejemplo, se puede indicar qué nodos tienen mayor valor en alguna de las medidas de centralidad y explicar por qué esos individuos tienen esos valores de acuerdo al significado de la red concreta que se está analizando. Así mismo, se puede hacer una discusión sobre el significado de las comunidades obtenidas. Como se van a analizar varias redes, la sección de discusión también ha de incluir información sobre las similitudes/diferencias que hay entre las distintas redes analizadas.

No se trata de escribir mucho ni de volver a explicar los datos de las secciones anteriores sino de hacer un análisis razonado de la red que estamos viendo, uniendo los conceptos vistos en clase con los conocimientos que tenemos de las redes analizadas.

4. Parte opcional

Se evaluará de manera muy positiva la realización de los informes para las dos opciones (correo y escenas de los Simpsons).

En este caso, os proporcionamos un conjunto de datos con información sobre los personajes que intervienen en las escenas de distintos episodios de Los Simpsons. Este conjunto de datos ha sido extraído de este conjunto de datos original contenido en Kaggle y tiene la siguiente información:

- `the_simpsons.csv`: Es un archivo de texto separado por comas. En cada línea aparece la siguiente información:
  - `Episode_id`: Identificador del episodio.
  - `Location_id`: Identificador de la escena.
  - `NamedCharacters`: Lista de los personajes que intervienen en la escena, separados por `:`. Solo aparecen las escenas en las que hay al menos 2 personajes.
- Adicionalmente, se proporcionan otros archivos CSV con información complementaria por si fuese necesaria para entender el contenido del anterior archivo.

Al igual que antes, es necesario crear un programa (en el lenguaje que se prefiera) para procesar los datos de este CSV y crear la lista de nodos y de aristas en un archivo en uno de los formatos que Gephi es capaz de cargar. Los nodos representarán cada uno de los personajes de la serie, mientras que una arista representa que esos dos personajes han aparecido juntos al menos en una escena. En este caso, deberemos crear un grafo *no dirigido* y es importante que las aristas tengan pesos (ya que representarán el número de escenas que comparten dos personajes). En general, Gephi es capaz de cargar aristas repetidas y computar el peso de la misma. Sin embargo, si esto no se produjese, entonces será necesario que vuestro programa calcule también el peso de cada arista.

5. Entrega

La práctica se entregará en el Campus Virtual, **antes de las 23:55 del día 4 de noviembre de 2018**.

La entrega de la práctica será un archivo `.zip` (etiquetado con el número de grupo *GrupoXX*) con los siguientes contenidos:

- *Memoria.pdf*: Un archivo pdf que deberá incluir, al menos, el siguiente contenido:
  - Portada con el número y título de la práctica.
  - Número de grupo.
  - Nombre y apellidos de los integrantes del grupo.
  - Una memoria con la información especificada en los apartados 2 y 3 del enunciado.
  - Referencias bibliográficas u otro tipo de material distinto del proporcionado en la asignatura que se haya consultado para realizar la práctica.
- Código de la aplicación desarrollada para crear los archivos en el formato para ser cargados en Gephi.
- Archivos de las redes generadas por la aplicación anterior y archivos de Gephi (`.gephi`) usados en el análisis.

El archivo puede ser subido por cualquiera de los integrantes del grupo (sólo una entrega).