Análisis de Redes Sociales

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

Curso 2015 - 2016

Prefacio

Estos son las prácticas de la asignatura Análisis de Redes Sociales, impartida en la Facultad de Informática de la Universidad Complutense de Madrid por los profesores Guillermo Jiménez Díaz y Alberto Díaz, del Departamento de Ingeniería del Software e Inteligencia Artificial.

Este material ha sido desarrollado a partir de distintas fuertes, destacando como referencia principal el libro *Network Science* de Laszlo Barabasi, el material de la asignatura *Social Network Analysis*, impartido por Lada Adamic a través de Coursera, y las transparencias de la asignatura Redes y Sistemas Complejos, creadas por Óscar Cordón García de la Universidad de Granada.

Práctica 1: Análisis básico de una red con Gephi

La práctica se dividirá en dos partes bien diferenciadas:

1. Procesado de datos: Se darán unos datos que no están en forma de grafo y que hay que procesar 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.

1.1 Procesado de los datos

Vamos a crear una red a partir de la información de películas y actores existente en la web http://www.linkedmdb.org/. Para evitar tener que hacer peticiones al servidor hemos creado un dataset formado por ficheros XML con información de unos 7700 actores y unas 5000 películas. La estructura del dataset (disponible en el Campus Virtual) es la siguiente:

- Directorio `movies`: Contiene un archivo XML para cada una de las películas del dataset, con nombre `data.linkedmdb.org.data.film.<id>.xml`, donde id es un identificador único de la película. Cada archivo tiene información sobre el título de la película que representa (contenido textual de la etiqueta `dc:title`) y de los actores que la interpretan (valor del atributo `rdf:resource` de la etiqueta `movie:actor`). Los actores se identifican con un identificador único que aparece al final del valor de dicho atributo y que se corresponde con uno de los ficheros contenidos en el directorio `actors`.
- Directorio `actors`: Contiene un archivo XML para cada uno de los actores del dataset, con nombre `data.linkedmdb.org.data.actor.<id>.xml`, donde id es un identificador único del actor. Cada archivo tiene información sobre el nombre del actor que representa (contenido textual de la etiqueta `movie:actor_name`), además de otra información adicional.

Primeramente es necesario a convertir la información del grafo bipartito, cuyas aristas se define mediante la relación `actor X actúa en la película Y` en un grafo *no bipartito*. Para ello hay que procesar los archivos del directorio `movies` y crear un grafo no bipartito cuyas aristas representan que `actor X y actor Y actúan en la misma película`. Para ello analizaremos cada uno de los XML y crearemos la lista de aristas con esta información. Además, podemos crear una lista de nodos con el identificador y el nombre de cada actor (extraído de los ficheros contenidos en el directorio `actors`). Una vez procesados todos los ficheros, exportaremos las listas de nodos y aristas creadas a un archivo en uno de los formatos que Gephi es capaz de cargar.

1.2 Visualización y extracción de información

Para las tareas de visualización, si la red presenta más de una componente conexa, se recomienda usar `Force Atlas 2` como algoritmo de layout (Distribución). Para evitar que las componentes conexas queden fuera de la vista principal que muestra la componente gigante, fijad el valor del parámetro `Puesta a punto >> Gravedad` en torno a 20. Si todo queda demasiado amontonado, se puede probar a marcar la opción `Disuadir Hubs` y/o `Evitar el solapamiento`. Los aspectos estéticos de la visualización se dejan al parecer del propio alumno, que puede probar las distintas variantes de algoritmos de layout implementados en Gephi y de parámetros para determinar cuál le proporciona la distribución que más le guste

Para los primeros pasos del análisis, comenzaremos por anotar los valores de las medidas globales básicas: número de nodos N y número de enlaces L que aparecen directamente en la ventana `Contexto`, además de calcular manualmente el número máximo de enlaces L_{max}. Posteriormente, calcularemos las restantes medidas globales (grado medio $\langle k \rangle$, diámetro d_{max} y distancia media $\langle d \rangle$) ejecutando las opciones correspondientes en la ventana `Estadísticas`.

Al realizar el cálculo del grado medio obtendremos también la distribución de grados de la red completa, que debemos grabar (Gephi lo guarda en una carpeta con una imagen png y un fichero html). El cálculo del diámetro nos proporciona también el valor de la distancia media, que anotaremos, y la distribución de distancias, que guardaremos, así como otras muchas medidas, varias de las cuales estudiaremos en temas de teoría posteriores como por ejemplo la Centralidad.

La opción Densidad de grafo nos mide la relación entre número de enlaces L y el número máximo de enlaces L_{max}. La ejecutaremos y anotaremos el valor. Finalmente, ejecutaremos la opción Coeficiente medio de clustering para obtener la medida del mismo nombre, $\langle C \rangle$. Dicha opción nos proporcionará también la distribución del coeficiente de clustering en la red, que guardaremos.

Ahora pasaremos a analizar la conectividad de la red. En primer lugar, obtendremos el número de componentes conexas ejecutando la opción `Componentes conexos` y lo anotaremos. Luego nos centraremos en la componente gigante y calcularemos su número de nodos. Para ello, iremos a `Filtros`, seleccionaremos `Topología >> Componente`

`gigante` y arrastramos el filtro a la ventana de abajo llamada `Consultas`, en donde pone `Arrastrar filtro aquí`. Entonces pulsaremos en el botón `Filtrar` con la flecha verde en la esquina inferior izquierda de la pantalla. La visualización cambiará y sólo mostrará la componente gigante. La ventana `Contexto` en la esquina superior izquierda nos mostrará el número de nodos y enlaces de dicha componente y sus porcentajes con respecto a la red total, los cuales anotaremos.

Una vez realizadas todo esto, cada alumno guardará el proyecto desde Gephi nombrándolo con sus apellidos y su nombre propio. Por otro lado, cada alumno del grupo almacenará todos los valores obtenidos en una tabla incluida en un fichero Excel llamado con el nombre del alumno.

La última tarea a realizar será escribir un pequeño análisis de las redes estudiadas a partir de los valores de medidas y de las gráficas de distribución de grados, distancias, etc. obtenidas. En particular hay que explicar como se ajustan los valores obtenidos a las propiedades definidas por el modelo de red aleatoria (al final del documento aparece un cuadro resumen de las principales propiedades). Hay que realizar un análisis individual de cada red, así como discutir la similitud o diferencias entre las redes analizadas por cada uno. No se trata de escribir mucho sino de hacer un análisis razonable considerando los conocimientos limitados que todavía tenemos sobre el análisis de redes.

Chapter 2

Entrega

La práctica se entregará en el Campus Virtual, antes de las 23:55 del día 2 de noviembre de 2014.

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

- *Documentación.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.
  - Para cada una de las redes de cada alumno
    - Una imagen de la red completa y otra de la componente gigante con una visualización lo más estética posible.
    - La tabla Excel con los valores de las medidas estudiadas.
    - Los gráficos de las distribuciones de grado, distancia, etc. generados por Gephi.
    - Un análisis de la red en función de los datos anteriores.
  - Una comparativa de las distintas redes analizadas.
  - Referencias bibliográficas u otro tipo de material distinto del proporcionado en la asignatura que se haya consultado para realizar la práctica.
- Ficheros obtenidos con netvizz para cada red.
- Ficheros de proyecto Gephi de los análisis realizados.
- Un fichero excel con los datos de todas las redes analizadas.

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

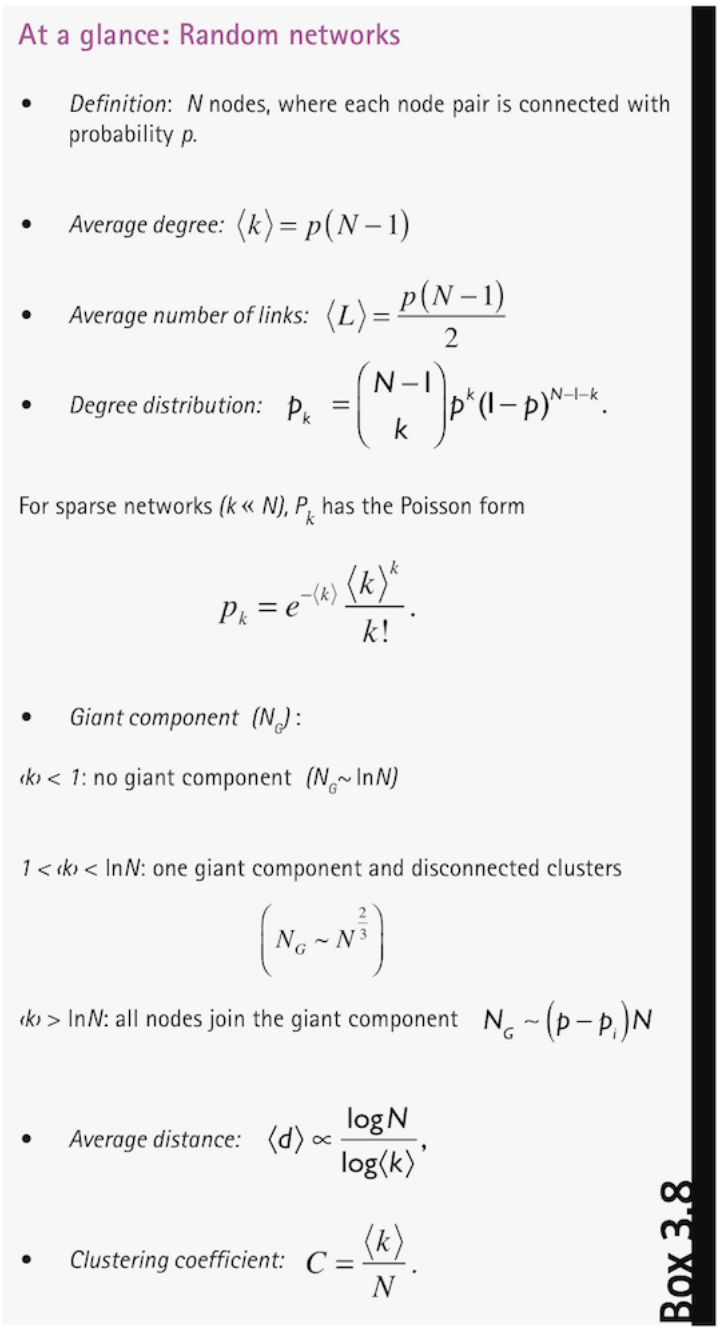

Figure 2.1: Resumen de las propiedades del modelo de red aleatoria