UNIVERSIDAD DE GRANADA
E.T.S.I. INFORMÁTICA Y TELECOMUNICACIÓN

Departamento de Ciencias de la
Computación e Inteligencia Artificial

Minería de Medios Sociales

Guión de Prácticas Bloque I.1:

Análisis y Visualización Básica de una Red Social con *Gephi*

Curso 2018-2019

Máster en Ciencia de Datos e Ingeniería de Computadores

Práctica Bloque I.1

Análisis y Visualización Básica de una Red Social con *Gephi*

1. Objetivos

El objetivo de esta práctica es doble. Por un lado, familiarizarse con los procedimientos de análisis de redes y con las medidas habitualmente consideradas para esta tarea. Por otro, aprender el manejo de una herramienta estándar de análisis y visualización de redes como *Gephi* [1], disponible para su descarga en https://gephi.org/users/download/.

Para ello, se requerirá que el estudiante escoja una red social, la cargue en la herramienta, la visualice y calcule los valores de una serie de medidas estándar de análisis de redes para estudiar las características principales de la misma así como la influencia de los distintos actores que la componen y su posible estructura de comunidades.

La práctica se evalúa sobre un total de **3 puntos**. La fecha límite de entrega será el **Lunes 13 de mayo de 2019** antes de las 23:59 horas. La entrega de la práctica se realizará en el espacio de la asignatura en la plataforma Prado.

2. Trabajo a Realizar

En esta práctica, la red a analizar será una red escogida de entre las disponibles en la literatura [2]. Estas páginas web incluyen repositorios de redes que se pueden emplear, así como cualquier otra propuesta por el estudiante y comunicada con anterioridad al profesor para su aceptación:

- Mark Newman: http://www-personal.umich.edu/~mejn/netdata/
- SNAP: http://snap.stanford.edu/data/
- UciNet: http://vlado.fmf.uni-lj.si/pub/networks/data/UciNet/UciData.htm
- Pajek: http://vlado.fmf.uni-lj.si/pub/networks/data/default.htm
- Gephi: https://github.com/gephi/gephi/wiki/Datasets
- Alex Arenas: http://deim.urv.cat/~alexandre.arenas/data/welcome.htm

[1] Aunque se recomienda el uso de *Gephi* y este guión de prácticas está personalizado para esa herramienta, el alumno puede optar por realizarla con cualquier otra de las herramientas de análisis y visualización de redes existentes.

[2] Como alternativa, también se permite que el estudiante analice una red obtenida de Twitter (o de cualquier otra red social electrónica) mediante el plugin de *Gephi* o usando cualquier otro *scraper*, que deberá ser comunicada y aceptada con anterioridad por el profesor.

El estudiante puede escoger la red que desee de entre las disponibles en estos repositorios o en cualquier otro pero deberá comunicarlo al profesor para que le confirme la asignación antes de comenzar a realizar la práctica. Esto se hace con objeto de que no se escojan las mismas redes en las prácticas de varios estudiantes. Para establecer una comunicación fluida en las dos direcciones (comunicación con el profesor y conocimiento de las redes que han escogido los compañeros), se ha habilitado un wiki en el espacio de Prado de la asignatura. **Se valorará positivamente el uso de redes con un tamaño razonable, al menos de unos pocos cientos de nodos**.

2.1. Análisis Básico de la Red

Una vez generada la red, se cargará en *Gephi* y se realizarán tareas básicas de análisis y visualización. Si la red presenta más de una componente conexa, se recomienda usar *Force Atlas 2* como algoritmo de *layout (*en la ventana *Distribución)*. Para evitar que las componentes conexas queden fuera de la vista principal que muestra la componente gigante, fijar el valor del parámetro *Gravedad* en *Puesta a punto* a un valor entre 10 y 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 estudiante, que puede probar las distintas variantes de algoritmos de *layout* implementados en *Gephi* y distintos valores 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 otra medida global, el grado medio $<k>$, ejecutando la opción correspondiente en la ventana *Estadísticas*. En el caso en que se nos preguntara, deberíamos especificar que la red es no dirigida. 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*).

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.

Posteriormente, ejecutaremos la opción *Coeficiente medio de clustering* para obtener la medida del mismo nombre, $<C>$. Dicha opción nos proporcionará también la distribución de coeficientes de clustering de la red, que guardaremos [3].

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* 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

[3] Hay veces que *Gephi* falla y devuelve una gráfica de coeficiente de clustering vacía. En ese caso, habrá que generarla a mano usando *Excel*. Para ello, basta con entrar en la pestaña *Laboratorio de datos* de *Gephi*, exportar los datos correspondientes en formato *csv* e importarlos en *Excel* para generar la gráfica correspondiente.

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.

Finalmente, calcularemos las restantes **medidas globales** (diámetro d_{max} y distancia media d) sobre la componente gigante de la red ejecutando la opción correspondiente al *Diámetro de la red* en la ventana *Estadísticas*. El cálculo del diámetro nos proporciona también el valor de la distancia media, que anotaremos, así como el de tres medidas de Centralidad (**intermediación**, **cercanía** y **excentricidad**), que emplearemos en la siguiente sección de la práctica.

La última tarea a realizar será escribir un pequeño análisis de la red estudiada a partir de los valores de medidas y de las gráficas de distribución de grados, etc. obtenidas. Será un análisis igual al que se realiza para las redes de proteínas de la levadura y de amistad de Facebook del profesor en las transparencias de la Sesión I.1 del curso. No se trata de escribir mucho sino de hacer un análisis razonable considerando los conocimientos limitados que tenemos sobre el análisis de redes.

2.2. Estudio de la Centralidad de los Actores

El estudiante realizará un pequeño análisis de redes sociales sobre la red basado en medidas de Centralidad. Determinará los 5 actores principales de la misma mediante las medidas de **grado**, **intermediación**, **cercanía** y **vector propio**.

El valor de tres de estas medidas ya está calculado con los pasos que hemos realizado en la sección anterior. La centralidad de grado (no normalizada) se generó al calcular el *Grado medio* en la ventana *Estadísticas*. Las de intermediación y cercanía se generaron con la opción *Diámetro de la red*. En este caso, sí que es posible especificar si se desean obtener normalizadas o no normalizadas con el *checkbox Normalizar centralidades en el rango [0,1]*. Finalmente, la *Centralidad de vector propio* se calcula en la opción del menú *Estadísticas* del mismo nombre.

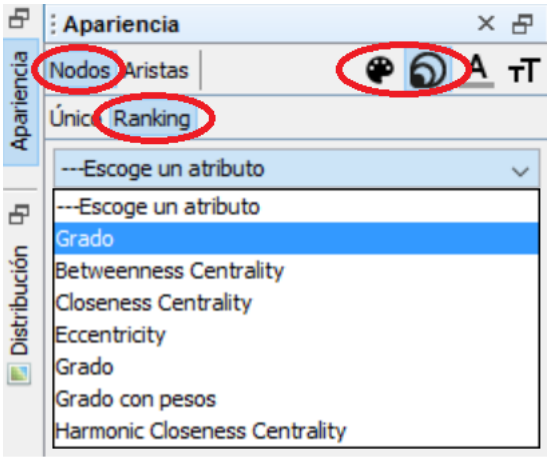
Los valores de centralidad de cada nodo pueden visualizarse en la tabla *Nodos* de la pestaña *Laboratorio de datos*, junto con el resto de la información asociada a cada nodo. Cada vez que se calcula una nueva medida usando las opciones de *Gephi*, aparece una nueva columna en esta tabla con sus valores. Se pueden ordenar los nodos por columnas simplemente pulsando sobre ellas. El estudiante anotará los nombres de los 5 actores con mejor valor para cada una de las cuatro medidas anteriores, así como el valor de dichas medidas y los almacenará en una tabla como la siguiente:

Centralidad de Grado	Centralidad de Intermediación	Centralidad de Cercanía	Centralidad de Vector propio
Nombre 1er actor: valor 1er actor	*Nombre 1er actor: valor 1er actor*	*Nombre 1er actor: valor 1er actor*	*Nombre 1er actor: valor 1er actor*
Nombre 2o actor: valor 2o actor	*Nombre 2o actor: valor 2o actor*	*Nombre 2o actor: valor 2o actor*	*Nombre 2o actor: valor 2o actor*
Nombre 3er actor: valor 3er actor	*Nombre 3er actor: valor 3er actor*	*Nombre 3er actor: valor 3er actor*	*Nombre 3er actor: valor 3er actor*
Nombre 4o actor: valor 4o actor	*Nombre 4o actor: valor 4o actor*	*Nombre 4o actor: valor 4o actor*	*Nombre 4o actor: valor 4o actor*
Nombre 5o actor: valor 5o actor	*Nombre 5o actor: valor 5o actor*	*Nombre 5o actor: valor 5o actor*	*Nombre 5o actor: valor 5o actor*

Finalmente, realizará un pequeño análisis de los actores más importantes de la red desde una perspectiva global en función de los valores de estas medidas y el conocimiento adquirido en la Sesión I.2 del curso.

Se valorará adicionalmente la realización de gráficas adicionales tales como:

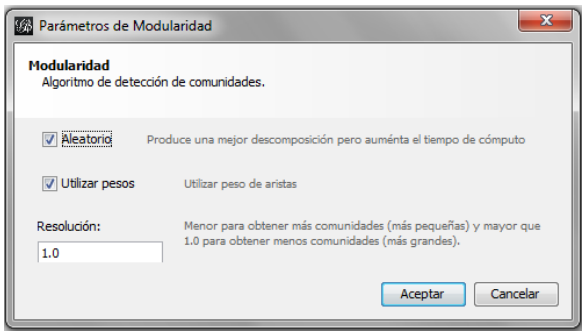
- Representaciones de la red en las que se visualicen dos de las medidas anteriores (por ejemplo, la intermediación en el tamaño de los nodos y la centralidad de vector propio en el color de los mismos) como las mostradas en las transparencias de la Sesión I.2 del curso. Estas visualizaciones pueden realizarse directamente en *Gephi*, usando las opciones *Nodos* y *Ranking* en la ventana *Apariencia*. Los dos iconos con la paleta y las bolas de distinto tamaño de la parte superior derecha de la pantalla permiten escoger qué valor de medida se desea emplear para definir el color y el tamaño de los nodos en la visualización, respectivamente:



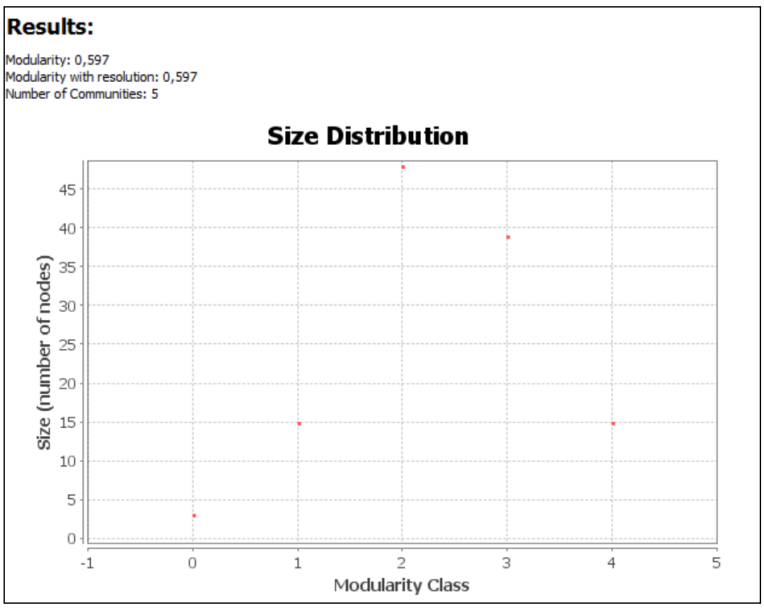
- Gráficos que representen los valores de dos de las medidas para todos los actores de la red en ejes de coordenadas como los estudiados en la Sesión I.2. Para realizarlos, puede exportar los valores de la Tabla de datos de la red en *csv* con la opción *Exportar tabla* y generarlos fácilmente usando Excel.

2.3. Detección de Comunidades

Se aplicará un método de detección de comunidades sobre la red estudiada para determinar la estructura modular de la red. Para ello, se usará el método de Lovaina, disponible en *Gephi*, ejecutando la opción *Modularidad* en la ventana *Estadísticas*:



El estudiante escogerá distintos valores para el parámetro *Resolución*, que determina el número de comunidades obtenido por el algoritmo, recordando que un valor más alto del parámetro genera un número menor de comunidades de mayor tamaño. Deberá perseguir la obtención de un número razonable que permita realizar un buen análisis de la estructura de comunidades obtenida. Mostrará los valores de la medida de modularidad asociados a cada particionamiento realizado y analizará la composición de las comunidades generadas para determinar si tienen algún tipo de influencia en la estructura de la red. Estos datos se muestran en la información que proporciona *Gephi* al ejecutar el método de Lovaina:



mientras que la composición de las comunidades en sí (la asignación de cada nodo a cada comunidad) pueden consultarse en la columna *Modularity Class* de la pestaña *Laboratorio de datos*. Realizará también dos o más visualizaciones de las particiones más significativas usando las opciones *Nodos* y *Partition*/*Ranking* en la ventana *Apariencia* para colorear los nodos en función de la comunidad a la que pertenezcan.

3. Documentación y Ficheros a Entregar

El estudiante guardará el proyecto desde *Gephi* nombrándolo con sus apellidos y su nombre propio. Luego almacenará todos los valores obtenidos en la tabla incluida en el fichero Excel disponible en el espacio de la asignatura en la plataforma, llamado *MedidasRedesPracticaMMS-I-1.xls*, renombrando el fichero de la misma forma.

La **documentación** de la práctica será un fichero *pdf* que deberá incluir, al menos, el siguiente contenido:

a) Portada con el título de la práctica, el curso académico y el nombre, DNI y dirección e-mail del estudiante.

b) Una sección que incluya:
   - 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 incrustada.
   - Los gráficos de las distribuciones de grado, etc.

c) Una sección que incluya el análisis de la red en función de los datos mostrados en la Sección 2.1.

d) Una sección que describa el análisis de la centralidad de los actores de la red desarrollado en la Sección 2.2.

e) Una sección que describa el estudio de las comunidades extraídas de la red en la Sección 2.3.

f) Una sección con las visualizaciones y gráficos adicionales (**en caso de haberlos realizado**).

g) Referencias bibliográficas u otro tipo de material distinto del proporcionado en la asignatura que se haya consultado para realizar la práctica (en caso de haberlo hecho).

Aunque lo esencial es el contenido, también debe cuidarse la presentación y la redacción.

El fichero *pdf* de la documentación, el fichero original de la red, el fichero del proyecto *Gephi* y el fichero Excel con los valores de las medidas se comprimirán conjuntamente en un fichero *zip* etiquetado con los apellidos y nombre del estudiante (Ej. Pérez Pérez Manuel.zip). Este fichero será entregado por internet a través de la plataforma PRADO2 (http://prado.ugr.es/).