Introducción a la visualización y análisis de redes con Gephi.

Mayo, 2020.

Introducción

Gephi es una herramienta muy popular para la visualización y anáisis de redes. Permite identificar propiedades ocultas, detectar patrones (comunidades) y hasta posibles errores en colecciones de datos.

En esta práctica, se realizarán análisis muy sencillos sobre algunas fuentes de datos previamente capturadas. Encontrará en comunidad.itam (o en el repositorio Test-Repo en github) los conjuntos de datos que utilizaremos en esta sesión.

Entregables

La práctica se puede hacer por parejas. La fecha límite de entrega es el viernes 19 de junio a las 17:00. Se entrega un documento con:

- Todas las preguntas solicitadas en este documento;
- Todas las capturas de grafos solicitadas;
- El código del script generado en la tercera parte

I. Visualización y análisis de redes en Facebook

Descargue el archivo `LadaFacebookAnon.gdf` desde el sitio del curso (carpeta `PracticaGephi`). Se trata de la red social anonimizada de la investigadora Lada Adamic [^Parte de esta sección esta basada en los trabajos y tutoriales de la investigadora Lada Adamic].

1. Abra la aplicación *Gephi* y seleccione `Nuevo proyecto`

2. En el menú archivo, presione `Abrir` y seleccione el archivo que acaba de descargar. Deje todas las opciones por omisión.

¿Cuántos nodos y cuántas aristas tiene su grafo?

La gráfica que se despliega no contiene información útil por ahora. Es un conjunto amorfo de nodos y aristas. Para poder interpretar mejor este grafo, es conveniente elegir una distribución apropiada.

Ayuda para visualizar su grafo

- Con la lupa que se encuentra en la ventana del grafo, puede centrar su red
- Con la rueda del ratón o bien con la barra zoom que encontrará expandiendo las opciones de la ventana del grafo (abajo, a la derecha), podrá acercar o alejar el diagrama de su red

- Mantenga presionado el botón derecho de su ratón para arrastrar el grafo y dejar en el área de visualización la sección de su interés

3. En el panel de `Distribución` (abajo a la izquierda) seleccione *Fruchteman Reingold*. Se trata de una distribución basada en los algoritmos de masa de partículas.

Ejecútelo con los parámetros por omisión y podrá observar cómo se van formando agrupamientos en el grafo. Cuando esté conforme con el resultado, detenga la ejecución del algoritmo.

Para esta sesión es más ilustrativo utilizar el desplegado *Force Atlas 2*. Está basado en un algoritmo de atracción y repulsión en función de la conectividad de los nodos: los nodos se repelen de acuerdo a un cierto criterio pero las aristas que los interconectan tratan de mantenerlos unidos. Como se mostrará más adelante, esta distribución es muy útil para identificar comunidades (clusters), mundos pequeños y de escala libre.

4. En el panel de `Distribución`, seleccione *Force Atlas 2*. Active la opción Evitar el solapamiento y ejecute nuevamente hasta que quede a gusto con el desplegado.

Capture la pantalla resultante (en el menú que se encuentra en la parte inferior de la ventana del grafo encontrará un ícono con la imagen de una cámara para hacer la captura de pantalla).

La visualización es más útil si los componentes del grafo tienen distintos colores. Para empezar, aprovechemos que una de las propiedades que tienen los nodos de esta red, es el género de sus miembros.

5. En el panel `Apariencia` (arriba a la izquierda), seleccione nodos y la opción *Atributo*. En el menú drop-down de atributos, seleccione sex y Aplicar. Observe cómo se colorea su grafo.

¿Qué porcentaje de hombres y mujeres tiene su grafo? ¿Cuenta con miembros que no reportaron su género? De ser así, ¿En qué porcentaje?

De no ser así, vaya al laboratorio de datos, elija algunos registros y elimine el género. Alternativamente, puede seleccionar un nodo y modificar sus propiedades. Aplique nuevamente el cambio de apariencia y reporte los resultados.

Además del color, Gephi permite desplegar algunas características del nodo (su tamaño, y el color y tamaño de las etiquetas) en función de algún criterio. En nuestro conjunto de datos, el único criterio disponible es el grado del nodo.

6. En el panel `Apariencia`, seleccione *Grado*, presione el ícono de tamaño [^Un grupo de círculos que se van agrandando. En las versiones anteriores, era un diamante], configure los parámetros Tamaño mínimo: 5 y máximo: 50, y aplique su configuración.

Si se enciman los nodos, vuelva a ejecutar la distribución *Force Atlas 2* con la opción de evitar solapamiento activada.

¿Cuáles son los cinco nodos que percibe que tengan el grado más alto? Puede seleccionar el nodo con la herramienta que ya conoce, o puede activar las etiquetas y ajustar su tamaño en la barra de opciones justo debajo de la ventana del grafo

7. Veamos qué comunidades puede detectar Gephi en la red. En la ventana `Estadísticas` identifique la opción *Modularidad* y ejecútela con una resolución de 1.5

¿Cuántas comunidades encontró?

8. Ahora coloree su grafo de acuerdo a las comunidades identificadas. Asegúrese de refrescar las opciones (abajo, a la izquierda) para que se despliegue y pueda elegir *Modularity class*

Capture la pantalla resultante

9. Analicemos ahora algunos de los datos que Gephi puede calcular desde la ventana `Estadísticas`

Ejecute *Grado medio*. **¿Cuál es el grado promedio obtenido?**

A partir del histograma que se despliega, la distribución se parece más a una uniforme, una normal o una power-law (por ejemplo, Pareto)

¿Qué tipo de red podría ser ésta: random, small-world, scale-free?

Ejecute *Diámetro de la red* **¿Cuál es el diámetro obtenido? ¿Qué significa esto? ¿Cuál es la longitud media de camino obtenida?**

¿Qué significa excentricidad? ¿Cuál fue la moda (valor con mayor nímero de ocurrencias) de excentricidad calculada?

Ejecute *Componentes conexos*. **¿Cuántos componentes fuertemente y cuántos débilmente conectados identificá?**

Podrá observar que hay nodos completamente aislados o con una conectividad sumamente baja. En muchas ocasiones es conveniente eliminarlos para poder concentrarse en los elementos más relevantes de la red.

10. En la ventana `Filtros` seleccione *Rango de grado* en Topología y arrástrelo al panel de consultas. Desplace la opción izquierda de la barra de configuración para que sólo se consideren nodos con un grado mayor a siete y presione *filtrar.*

¿Cuántos nodos y cuántas aristas tiene ahora su grafo?

¿Cuáles son los nuevos valores de grado medio, diámetro de la red y modularidad? Para el último valor, mantenga la resolución de 1.5. ¿Cuántas comunidades encontró?

II. Visualización de flujos de bienes

En esta sección [^Esta sección esta basada en los tutoriales de Martin Grandjean] vamos a utilizar los archivos `Nodes1.csv` y `Edges1.csv`. Contienen información de cartas enviadas entre distintos países de Europa. Descárguelos desde el repositorio del curso (carpeta `PracticaGephi`).

1. Cree un nuevo proyecto. Si desea, puede guardar el anterior. Utilizaremos dos complementos (**plugins**): `GeoLayout` y `MapOfCountries`. Se encuentran en el menú de `Herramientas`.

Ábralo, busque el complemento y selecciónelo. Tendrá que reiniciar Gephi nuevamente.

2. En el `Laboratorio de Datos` seleccione *Importar hoja de cálculo* y cargue el archivo `Nodes1.csv`.

En las opciones, especifique que la separación entre columnas es el caracter punto y coma y que se está importando la tabla de nodos.

Seleccione *siguiente* y asegúrese que los *parámetros de importación* son correctos. En particular, revise que los argumentos de longitud y latitud sean tratados como *double* y no como *string* o *integer*.

3. Ahora importe el archivo `Edges1.csv` como *tabla de aristas* con los parámetros apropiados. Deseleccione "crear nodos inexistentes" porque ya cargó los nodos.

Visualización

Ahora cámbiese al panel de `Vista general`. Podrá comprobar que, como en la sección anterior, la gráfica desplegada simplemente no se puede entender. La iremos enriqueciendo poco a poco.

4. Empecemos por asignar a los nodos un tamaño proporcional a su grado (sin distinguir entre grado de entrada o de salida). Recuerde que esto se hace en el menú de `Apariencia/Nodo/Atributos`. Asigne valores mínimo y máximo de 10 y 100 respectivamente.

En la opción *spline* puede modificar la relación de grado a volumen si desea que ésta no sea linealmente proporcional. A veces se busca ese efecto para enfatizar las diferencias de grado: Recordemos que nuestra percepción visual es frágil para distinguir volúmenes.

Capture la pantalla resultante

5. Ahora trabajemos con la distribución. Para espaciar los nodos pero mantenerlos dentro de un área acotada, utilice de nuevo *Fruchterman Reingold* con valores de 20000, 10 y 10 para el área, la gravedad y la velocidad respectivamente.

Si lo ha hecho bien, puede empezar a distinguir algunas comunidades en la red, pero una vez más conviene afinar el grafo con el algoritmo *Force Atlas 2* para dispersar los grupos. Seleccione *Evitar el solapamiento* y cambie el *Escalamiento* a 50. Deje correr el algoritmo hasta que el grafo está relativamente estable y deténgalo.

Capture la pantalla resultante

Mediciones de centralidad

Al hacer algunos cálculos de estadísticas, se agregarán atributos a los nodos que nos permitirán incrementar la calidad informativa del grafo.

6. En el `Laboratorio de Datos` seleccione la tabla de aristas y ordénela de acuerdo a su peso (de clic sobre la etiqueta *weight*). Como podrá observar, el peso de las aristas va de 1 a 3. Por ello, debemos tomar en cuenta esas diferencias al calcular el grado de los nodos. Debe hacerse un cálculo de grado ponderado.

También habrá notado que se trata de un grafo dirigido: Las aristas tienen una dirección de la fuente al destino como se muestra por las flechas con que se representan las aristas. Por ello, al calcular el grado se deben tomar en cuenta las conexiones de entrada y de salida.

En el panel de `Estadísticas` de clic en *Ejecutar Grado medio con pesos*

¿Cuál es el valor de grado medio con pesos?

7. Ahora que se han calculado estos valores, podemos usarlos para discriminar los nodos en función de su grado. En el panel de `Apariencia/Nodos/Ranking`, seleccione el ícono de color y el atributo *Grado de entrada con pesos* para discriminar los nodos con base en el número de aristas de entrada.

Sugerencia: Invierta la barra de colores para usar un color claro para los nodos con una conectividad alta y uno oscuro para valores pequeños de manera que estos últimos se mantengan visibles en el grafo.

Capture la pantalla resultante

Como puede observar, los nodos más grandes (los que tienen el grado más alto) no necesariamente tienen también el grado ponderado de entrada más alto (no necesariamente son los más claros): Los nodos que generan más flujos de salida no necesariamente son los que reciben más.

En conclusión, resulta útil asignar distintos atributos al tamaño y el color de los nodos para poder compararlos.

8. En el panel para editar las propiedades de las etiquetas de los nodos (fuente, color y tamaño), seleccione *Etiquetas* y de clic en *nodos*.

Al hacer zoom en su grafo, podrá ver que los nodos ya aparecen con un nombre. En el área del panel tiene una barra deslizadora para modificar el tamaño de las etiquetas.

9. Para resaltar las comunidades en esta red, las cuales dependen de una comparación entre las densidades de aristas en un grupo y del grupo con el resto de la red, calculemos la *Modularidad* en el panel de `Estadísticas` con una resolución de 0.8. Como lo hemos hecho anteriormente, en el panel `Apariencia/Nodos/ParticionAtributos`, seleccione *ModularityClass* para resaltar las comunidades.

Capture la pantalla resultante

10. La longitud media del camino o centralidad intermedia (*betweenness centrality*) mide todas las distancias más cortas entre todos los pares de nodos en la red y después cuenta cuántas veces un nodo está en la ruta más corta entre otros dos. Puede ser una métrica interesante para identificar nodos que parecen fungir como intermediarios entre dos entidades.

De clic en *Diámetro de la red* y *grafo dirigido*. Observe que tanto el diámetro como la longitud media del camino están siendo calculados. Esto puede tomar algún tiempo.

Muestre diámetro y longitud media del camino

Como en el caso anterior, seleccione una paleta de colores atractiva para resaltar nodos en función de su centralidad intermedia. Es bastante claro que nodos con un alto grado no tienen una centralidad intermedia alta.

Terminando el grafo

Seleccione la sección `Previsualización` para realizar unos últimos ajustes. A diferencia de las configuraciones anteriores, los cambios en esta sección son reversibles y no afectan la estructura del grafo.

Puede jugar con los parámetros de configuración. Se sugiere fijar la opacidad de las aristas a 70% para obtener un buen contraste entre los nodos.

Las aristas curvas son una convención en grafos dirigidos y su dirección se interpreta en el sentido de las manecillas del reloj. Las aristas rectas suelen reservarse para grafos no dirigidos.

Juegue con los parámetros y cuando está contento con el resultado, guarde su grafo en formato svg y en pdf.

Distribución geográfica

Al importar el archivo csv pudo observar que nuestros datos tienen coordenadas geográficas. El componente *Geo Layout* nos permite desplegar los nodos con base en estas coordendas.

11. En el panel de `Distribución`, seleccione *Geo Layout* y asigne una escala de 20000. Asegúrese que *Latitude* y *Longitude* se corresponden con los atributos de nuestros datos y

que la proyección es *Mercator* (para poder ser incorporada al mapa que usaremos posteriormente). Quizás tenga que cambiar de espacio de trabajo.

Como los nodos están agrupados por coordenadas geográficas, utilice la distribución *Noverlap* para separarlos dentro de cada coordenada.

Quizás pueda tener una vista más agradable si suaviza el grosor de las aristas en el panel inferior del área de visualización.

Aunque la distribución *MapOfCountries* es muy pobre, nos da una aproximación de la ubicación de los nodos cuando éstos tienen coordenadas geográficas.

12. Seleccione la distribución *MapOfCountries*, la región *Europe* y asegúrese que está activada la proyección *Mercator*.
13. En el panel de previsualización modifique a su gusto la apariencia final y guarde el resultado como un archivo .svg y .pdf

Muestre el resultado

NOTAS

```
* Problemas con los controladores por una actualización que se está haciendo.
* Deshabilitar controlador gráfico mientras se utiliza Gephi (Intel Graphics) y
habilitarlo nuevamente al terminar (https://github.com/gephi/gephi/issues/2126)
```