TP1: Data mining en Grafos

Sebastián A Romano, Hernán Varela, Santiago Bezchinsky, Julián Devouassoux
Data Mining en Ciencia y Tecnología

14 de abril de 2025

1. Objetivos

El objetivo principal de este trabajo será aplicar los conceptos que se discutieron en las clases de Grafos, utilizando un enfoque científico: plantearse una pregunta e investigar su posible respuesta. Les vamos a dar un conjunto de datasets de grafos y ustedes podrán elegir cuáles de ellos van a analizar (según las consignas). Esta elección debe estar motivada por una pregunta inicial que servirá de origen a todo el análisis subsiguiente. Dicha pregunta debe estar planteada y justificada en la sección *Introducción* (idealmente con bibliografía) y discutida en la sección *Discusión* de acuerdo a los resultados obtenidos. Por ejemplo, esta pregunta puede ser del tipo: "por X motivo esperamos que el dataset A tenga la propiedad Y más exacerbada que el dataset B, y nos proponemos investigar si esto es así". Con esta pregunta como guía van a realizar una variedad de análisis que punteamos más abajo, y comparar los resultados obtenidos en los datasets elegidos.

2. Información sobre los datos

Los datasets que les proponemos son:

- **Los Miserables**: Una red no dirgida y pesada de co-ocurrencias de personajes en la novela de Vitor Hugo "Los Miserables". Un nodo representa un personaje y un enlace entre nodos indica que ambos personajes aparecen en el mismo capítulo de la novela. El peso de cada enlace indica cuantas veces co-ocurren. Contiene 77 nodos y 254 enlances.

- **Terroristas**: Una red no dirigda y pesada que representa la red de contactos entre supuestos terroristas envueltos en el atentado a un tren en Madrid el 11 de Marzo del 2004, reconstruida a partir de las noticias en los diarios. Un nodo representa un terrorista, un enlace indica un contacto entre terroristas, y el peso indica cantidad de contactos. Contiene 64 nodos y 243 enlances.

- ***C. elegans***: Una red no dirigida y pesada que representa la red de conexiones neuronales en el organismo *Caenorhabditis elegans* (un nematodo). Cada nodo es una neurona, y cada enlace una conexión entre neuronas (en el dataset original los enlaces eran

direccionales, aquí la direccionalidad se ha omitido). El sistema nervioso de este organismo es el único al cual se le han podido reconstruir todas sus conexiones neuronales. Contiene 297 nodos y 2148 enlances.

- **Facebook**: Una red no dirigida y no pesada de círculos de Facebook. Los datos fueron regoidos por una encuesta realizada en la app de Facebook. Cada nodo es un perfil de Facebook y un enlace significa que los perfiles son amigos en Facebook. Contiene 4039 nodos y 88234 enlaces.

- **Email-EU**: Una red no dirigida y no pesada de contactos vía e-mail en una institución europea de investigación. Cada nodo es una persona dentro de la institución, un enlace significa que estas personas intercambiaron al menos un email. Para este dataset también tenemos un label conocido de antemano para cada nodo, que representa a qué departamento de la institución pertenece el nodo. Contiene 1005 nodos y 25571 enlaces.

- **Aeropuertos**: Una red dirigida y pesada. Cada nodo es un aeropuerto de Estados Unidos, y los enlaces representan vuelos realizados en el año 2010. El peso de cada enlace indica el número total de vuelos de dicha ruta. Contiene 1574 nodos y 28236 enlaces.

Ejemplos de los procedimientos para la descarga de los datasets, así como varias funciones útiles para comenzar el análisis pueden encontrarse acá.

3. Comentarios generales

Las consignas que les planteamos en este TP incluyen comparación con modelos prototipo, análisis de centralidad de los nodos, robustez de los grafos y detección de comunidades. Con estas herramientas les proponemos que comparen entre los datasets que vayan a elegir. Esto es sólo una hoja de ruta. Cualquier otro análisis que propongan ustedes es más que bienvenido.

Deberán elegir al menos 2 datasets, pero no más de 3. Para motivar la pregunta que determine la elección de los datasets, les sugerimos que de forma preliminar exploren los datos: visualicen el grafo, calculen algunas variables topológicas que los caractericen (grado medio, coeficiente de clustering medio, distribución de grado, etc...) y vean si hay alguna diferencia entre los datasets que les gustaría explorar.

Recomendaciones: Para grafos no dirigidos, si el grafo no es conectado analizar la componente gigante únicamente. Para grafos dirigidos analizar solo la componente fuertemente conectada. En ambos casos, eliminar autoenlaces (enlaces que nodos realicen a sí mismos), si los hubiere.

4. Consignas

(a) **Descripción de los datasets elegidos.** Describir las características generales de cada dataset, así como sus variables topológicas que les parezcan relevantes y que hayan motivado la pregunta que se plantean.

(b) **Comparación con prototipos.** En este punto trabajar solo con grafos no dirigidos ni pesados. Asi que si eligieron un grafo que no cumpla con estas características, transformalo a un grafo no dirigido ni pesado para este punto en particular. Comparar los datasets con prototipos equivalentes de *Erdos-Renyi*, *Watts-Strogatz* y *Barabasi-Albert*. Explicar el interés de utilizar estos prototipos para la pregunta en cuestión. Sugerimos que comparen la distribución de grado de los datasets con la de una instancia de los prototipos, así como comparar el coeficiente de clustering medio y la distancia mínima media de los datasets con la de varias instancias de los prototipos. Discutir similitudes, diferencias en las comparaciones, validez/utilidad de los prototipos para modelar sus datasets. ¡Sean críticos de los resultados!

Opcional: si el modelo de *Barabasi-Albert* no es apropiado (porque resulta en redes con bajo clustering), intentar con el modelo de *Holme-Kim*.

(c) **Análisis de centralidad de los nodos.** Visualizar grafos de los datasets elegidos indicando la centralidad de cada nodo (por ejemplo, coloreando y/o escaleando el tamaño del nodo). Para grafos no dirigidos sugerimos al menos comparar la centralidad de intermediación y la de grado, mientras que para grafos dirigidos al menos comparar centralidad de grado y de PageRank (pero pueden usar otras centralidades si les resulta útil para explorar su pregunta). Discutir la información que cada una de estas medidas de centralidad proporcionan. ¿Son complementarias? ¿Redundantes? ¿Qué *insight* se puede obtener con estas medidas?

(d) **Análisis de robustez de los grafos.** Hacer un análisis de robustez, en el cual se borran los enlaces de un nodo a la vez, hasta borrar todos los enlaces del grafo. Comparar el efecto de borrar seleccionando nodos al azar, con el efecto de ir borrando primero los enlaces de los nodos con mayor centralidad. Usar las medidas de centralidad que se calcularon en el punto anterior. A medida que se van borrando los enlaces, para grafos no dirigidos cuantificar la robustez grafos por dos vías: 1) calculando el tamaño relativo de la componente gigante resultante (N_g/N) ; 2) calculando la *eficiencia global* del grafo. Para grafos dirigidos usar solo N_g/N (ya que la eficiencia global no está implementada para grafos dirigidos).

(e) **Comunidades en los grafos** Detectar la partición óptima (según la *modularidad*) utilizando el algoritmo de *Louvain*. Reportar el valor de *modularidad* obtenido. Visualizar las comunidades en el grafo y evaluar cualitativamente el resultado. Comparar entre los datasets analizados. Probar también con el algoritmo de *Girvan-Newman* y comparar resultados. Si eligieron el dataset Email-EU, comparar las comunidades obtenidas con los labels que indican el departamento de cada nodo, discutiendo similaridades y diferencias. ¿Cómo se comparan las comunidades encontradas con los labels dados por la pertenencia de los nodos a los distintos departamentos?

Formato

Les proponemos seguir el formato de publicación en una revista científica, pueden encontrar muchos formatos directamente en *Overleaf*, por ejemplo *NeurIPS*[1] o *IEEE Conference Template for ANCS 2019*. No es obligatorio seguir ese formato, pero si elegir el formato de alguna revista.

Las revistas suelen tener además instrucciones respecto al formato online, desde restricciones en el tamaño de cada sección, en el número de figuras/tablas, las secciones que debe contener, hasta formato de los números, referencias, etc.

Aquí ponemos nuestras restricciones, pero si quieren adaptarlo a alguna revista en particular también vale (explicítenlo en el informe).

Secciones.

1. Título (máx. 100 caracteres), tiene que ser expresivo (no vale TP1).

2. Resumen (máx. 200 palabras), tiene que contener una descripción de todo el trabajo: Motivación, Antecedentes, Objetivos, Métodos, Resultados y alguna Conclusión.

3. Introducción Comienza con la motivación, sigue con los antecedentes, y termina siempre con un párrafo de objetivos (no es necesario que este dividido en sub-secciones). Típicamente, una vez que motivaron el trabajo y mostraron lo que hay hecho, viene una frase del estilo "Por ende, nos proponemos..." o "Aquí nos proponemos...".

4. Métodos Detalle de los métodos a utilizar, en este caso no es necesario profundizar mucho pero pueden enumerarlos y sobre todo es el lugar para incluir cualquier método fuera de lo común que hayan utilizado.

5. Resultados y discusión Aquí se enumeran y discuten los resultados. Es muy importante que no sea una seguidilla de figuras y tablas. Como regla pueden considerar: *"Si una figura no se describe/comenta en el texto es que: O bien está de más y no hace a la historia, o bien se olvidaron de incluirla."*.

6. Conclusiones Comienza generalmente con un resumen muy breve de los principales resultados obtenidos (uniendo distintas secciones), y luego se pasa a conclusiones generales, detallando problemas detectados, posibles explicaciones y trabajo a futuro.

7. Referencias Citas bibliográficas utilizadas durante el reporte. Si son sitios web o repositorios se incluyen generalmente al pie de la página que corresponde y no como cita bibliográfica.

Considerando Introducción, Métodos, Resultados y Conclusiones no deben superar las 5000 palabras (aprox.). Finalmente, considerando el formato de este trabajo pueden dividirlo en

1. Título

2. Resumen

[1] https://www.overleaf.com/read/mzbjfyxsfxqn

3. Introducción
4. Métodos generales (si los hubiese)
5. Resultados
6. Conclusiones
7. Referencias

Figuras y tablas

Es muy importante pensar y tomar la decisión de qué mostrar y cómo mostrarlo. A veces no se le da importancia cuando el espacio no está acotado pero igualmente afecta seriamente a la comunicación de los resultados, ya que una mala visualización puede esconder lo relevante y lo que se quiere comunicar. Les dejamos algunas referencias para explorar al respecto especificamente de las decisiones, en cuanto a cómo implementarlo podemos discutirlo en clase [1, 2, 3]

Comentarios finales

Se puede usar cualquier herramienta de análisis o combinación de herramientas, debiendo indicarlas en el informe. Si usan una función ya armada dentro de una librería, detallen los parámetros con la que la corrieron. El lenguaje (Python o R) en el que se desarrolle el TP no es excluyente.

Referencias

[1] Seán I O'Donoghue, Benedetta Frida Baldi, Susan J Clark, Aaron E Darling, James M Hogan, Sandeep Kaur, Lena Maier-Hein, Davis J McCarthy, William J Moore, Esther Stenau, et al. Visualization of biomedical data. *Annual Review of Biomedical Data Science*, 1(1):275–304, 2018.

[2] Stephen R Midway. Principles of effective data visualization. *Patterns*, 1(9):100141, 2020.

[3] Elena A Allen, Erik B Erhardt, and Vince D Calhoun. Data visualization in the neurosciences: overcoming the curse of dimensionality. *Neuron*, 74(4):603–608, 2012.