Universidad de Castilla-La Mancha

Álgebra y Matemática Discreta

Practica Incremental 2022-23 (Grupo Español)

Visualización de la estructura de grupos de investigación

Profesores de la asignatura:
José Ángel Martín Baos [Grupo A] (joseangel.martin@uclm.es)
José Luis Espinosa Aranda [Prácticas de los Grupos C y D]
(josel.espinosa@uclm.es)
Ricardo García Ródenas [Grupo B] (ricardo.garcia@uclm.es)
José Ángel López Mateos [Teoría de los Grupos C y D]

Agradecimientos especiales:
Jared Guerrero Sosa.

A tener en cuenta:

1. La realización de la práctica es individual.
2. Esta práctica se evalúa sobre 2.5 puntos del global de la asignatura y vale aproximadamente un $70\,\%$ de la parte de prácticas de la asignatura.
3. La práctica es una actividad obligatoria, por lo que será necesario obtener al menos un $40\,\%$ de la puntuación (1 punto) para poder aprobar la asignatura.
4. No es necesario realizar todos los hitos para que la práctica pueda ser entregada.
5. Para la evaluación de cada uno de los apartados se tendrán en cuenta tanto la claridad del código como los comentarios incluidos en éste.
6. Para la corrección de la práctica se utilizará un programa de detección de copia. En caso de que la similitud entre dos o más prácticas se encuentren fuera de lo permisible, todas ellas serán calificadas con un 0 y aparecerá suspenso en la asignatura tanto en la convocatoria ordinaria como extraordinaria.
7. Fecha límite de entrega: Hasta el 21 de Mayo. Se habilitará una tarea en Campus Virtual para poder subirla.
8. Los códigos se entregaran en formato script de MATLAB (.m). No se aceptarán otros formatos, como por ejemplo cuadernos de MATLAB (.mlx).
9. Los códigos deben entregarse comprimidos en un único fichero .zip con el nombre del alumno. No olvides adjuntar todos los ficheros necesarios para su correcta ejecución, incluidos los ficheros proporcionados en Campus Virtual.

Objetivos de la práctica:

El objetivo de esta práctica es aplicar la teoría de grafos para modelar un problema bibliométrico en el contexto de la investigación llevada a cabo por la Universidad de Castilla-La Mancha (UCLM). A través del uso de técnicas de análisis de grafos, se pueden identificar patrones y relaciones entre los investigadores y sus publicaciones, así como obtener una representación gráfica de la estructura de los grupos de investigación establecidos. Esto puede ayudar a entender mejor la dinámica de colaboración y el impacto científico del grupo. Además, también se pueden utilizar estos resultados para identificar posibles áreas de mejora y oportunidades de colaboración futura entre investigadores.

En concreto, en esta práctica se busca resolver un primer problema más sencillo, el de encontrar automáticamente la *estructura* de los miembros de un grupo de investigación. Para ello, se trabajará con un fichero de datos que contiene las publicaciones científicas de los 12 investigadores de la UCLM con el mayor índice h (*h-index*) y sus colaboradores, recogidas hasta Enero de 2023. El índice h es un sistema de medición de la productividad y el impacto de un investigador. Este índice se basa en el número de publicaciones científicas de un investigador y el número de veces que esas publicaciones han sido citadas por otros investigadores. El valor de h se refiere al número más alto de publicaciones de un investigador que han sido citadas al menos h veces cada una. Por ejemplo, si un investigador tiene un índice h de 10, significa que tiene 10 publicaciones que han sido citadas al menos 10 veces cada una.

Los datos anteriores han sido recogidos utilizando la plataforma Scopus. Scopus es una base de datos bibliográfica propiedad de Elsevier que contiene más de 70 millones de registros de artículos de investigación, libros, tesis, conferencias y otros documentos científicos. Scopus se utiliza ampliamente en el ámbito académico y científico para la recuperación de información, la medición del impacto, la detección de tendencias y la identificación de oportunidades de colaboración.

Para ello se han obtenido tres ficheros de datos proporcionados en la carpeta `Data/`:

- `1_authors.csv`. Este fichero contiene una tabla con los datos de los distintos autores de la UCLM que han publicado conjuntamente con alguno de los 12 investigadores con mayor índice h. Esta tabla contiene las siguientes columnas con el siguiente tipo de datos:
  - `ID_author` (entero de 32 bits). Contiene un índice que identifica unívocamente al autor.
  - `AU_ID_scopus` (entero de 64 bits). Representa el ID del perfil del autor en la plataforma Scopus, de la cual se han obtenido los datos.
  - `full_name` (cadena de texto). Nombre completo del autor tal y como aparece reflejado en Scopus.
  - `affiliation` (cadena de texto). Universidad o centro de investigación en el que trabaja el autor.
- `2_publications.csv`. Este fichero contiene los datos de las publicaciones obtenidas de Scopus. **No es necesario para resolver esta práctica**, y por lo tanto, se proporciona únicamente para mostrar al alumno los datos completos.
  - `scopus_id` (entero de 64 bits). Contiene un índice que identifica la publicación únivocamente en Scopus.
  - `document_type` (cadena de texto). Tipo de contribución.
  - `doi` (cadena de texto). El DOI (Digital Object Identifier) es una cadena de caracteres única asignada para identificar en linea contenidos científicos y académicos como artículos, libros electrónicos, tesis y otros tipos de documentos.
  - `source_type` (cadena de texto). Tipo soporte de la publicación.
  - `source_title` (cadena de texto). Nombre de la revista.
  - `document_title` (cadena de texto). Título de la publicación.
  - `publication_date` (fecha). Fecha de la publicación.
  - `citations` (entero de 32 bits). Número de citas a fecha de obtención de los datos.
- `3_collaborations.csv`. Fichero que contiene las colaboraciones realizadas entre pares de autores hasta Enero de 2023. Se entiende por colaboración la publicación como coautores de un artículo, libro o actas de congreso científico. Los datos contenidos por este fichero son:
  - `ID_author_1` (entero de 32 bits). Índice del primer autor.
  - `ID_author_2` (entero de 32 bits). Índice del segundo autor.

- `scopus_id_collaborations` (cadena de texto). Lista con los identificadores de las publicaciones en co-autoria por el par de autores, separadas por comas.

Con estos datos como punto de partida, se pide que se lleven a cabo los siguientes hitos de manera individual, utilizando para ello el lenguaje MATLAB y los conceptos estudiados en las sesiones de prácticas y teoría.

Hito 1: Construcción del grafo de colaboración (0.5 puntos)

En este hito se propone construir un grafo para representar las relaciones entre los distintos investigadores de la UCLM. Los vértices del grafo están asociados a los autores y las aristas representan la colaboración entre dos de estos autores. En concreto, dos investigadores están conectados (por una arista) y solo si tienen al menos un trabajo en común.

Se pide implementar un script de MATLAB llamado `Milestone1.m` que realice los siguientes apartados:

a) Carga los ficheros necesarios utilizando los tipos de datos adecuados. Se proporciona una plantilla denominada `Milestone1Template.m` que debes usar para resolver este hito.

b) El peso de las aristas del grafo que se va a construir debe medir la intensidad en la colaboración entre un par de autores. Para calcular el peso de cada arista se podría pensar en sumar el número de publicaciones. Esta opción tiene el inconveniente que no todos los trabajos contribuyen de la misma forma. Esto es, que un trabajo con muchos autores aumentaría en mayor medida el valor de las aristas. Para corregir este efecto, se va a definir la aportación de una publicación como

$$1/\text{número de pares de autores en la publicación}$$

Por ejemplo si un artículo está firmado por los autores A, B, C entonces tendremos en nuestro grafo (fichero `3_collaborations.csv`) las aristas $A-B$, $A-C$ y $B-C$, por lo que el trabajo contribuye a estas tres aristas con un peso de $\frac{1}{3}$. El peso total de la arista será la suma de todos los pesos de los trabajos que tiene dicha arista.
Nota: Observa que estamos ante un grafo no dirigido, por lo que el orden de los vértices al definir las aristas no es relevante.

c) Construye el grafo anterior. Para ello, utiliza la tabla de autores obtenida del fichero `1_authors.csv` como tabla de nodos. Para utilizar esta tabla como tabla de nodos se deben realizar dos cambios en la tabla: primero se debe renombrar el campo `ID_author` a `Name` y segundo se debe hacer un cambio del tipo de dato de esta columna a `string`.
Nota: Consulta la documentación de MATLAB para ver como crear un grafo a partir de una lista de nodos de origen, otra de nodos de destino, un vector de pesos y la tabla de nodos.

d) Representa el grafo en MATLAB. Para ello, crea una figura denominada: "Collaboration graph". Investiga y juega con los parámetros de la visualización hasta conseguir un resultado elegante. Por ejemplo, prueba a utiliza los distintos algoritmos proporcionados

en el parámetro 'Layout' del comando `plot`, o a modificar otros parámetros como el tamaño de marcadores, líneas, colores, formas, etc. La Figura 1 muestra un ejemplo de una posible representación válida.

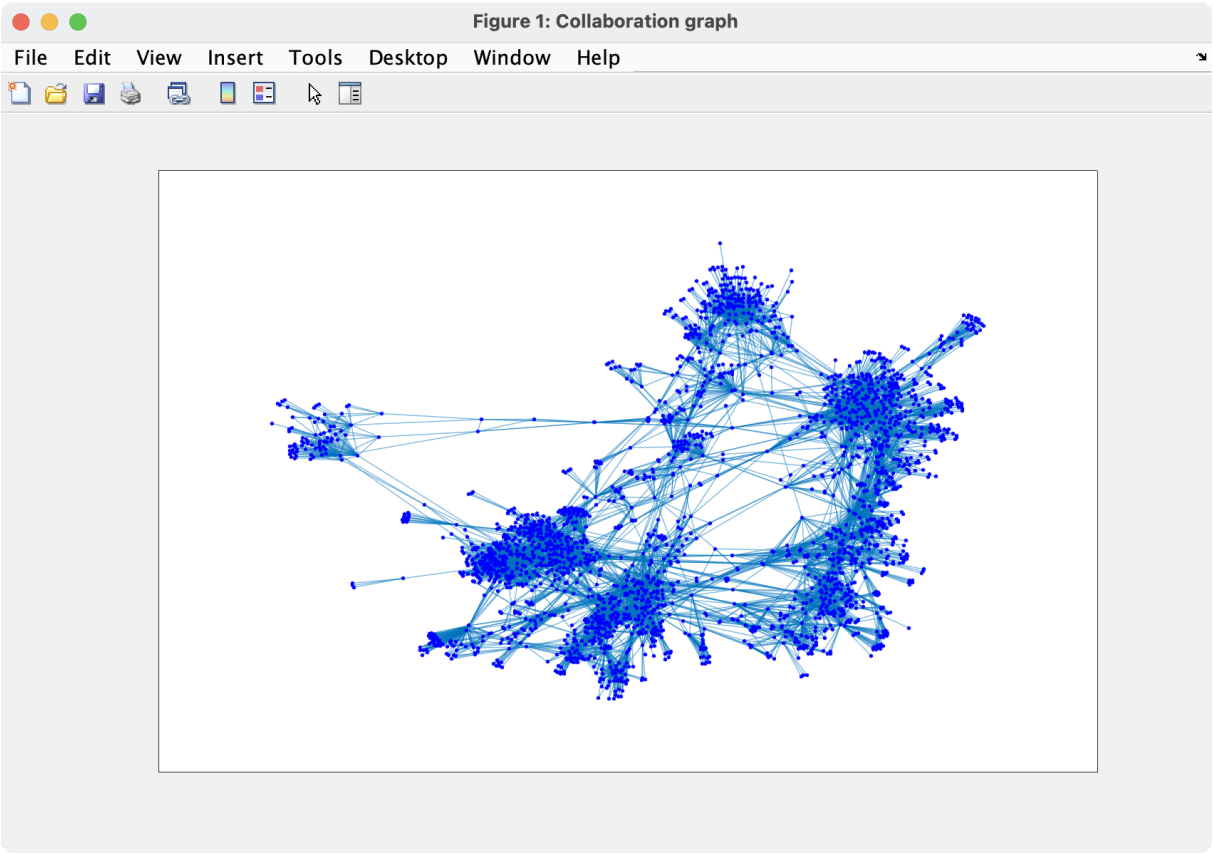

Figura 1: Ejemplo de representación del grafo obtenido en el Hito 1

Hito 2: Determinar los grupos de investigación (0.75 puntos)

Un grupo de colaboración consta de las relaciones internas de uno o varios grupos de investigación y las externas con otros miembros de la comunidad investigadora. Por lo tanto, los grupos de investigación son subgrupos del grafo de colaboración. Los miembros de un grupo de investigación colaborarán de manera más frecuente, lo que implica que los pesos de sus aristas serán mayores, creando comunidades con relaciones más fuertes en el grafo. En este hito se busca determinar automáticamente distintos grupos de investigación en el grafo de colaboración.

El procedimiento propuesto consiste en eliminar las aristas cuyo coste sea menor que una cantidad dada o umbral, α. Esto creará varias componentes conexas en el grafo, es probable que cada componente conexa componga un grupo de investigación. El problema es elegir automáticamente el umbral α más adecuado. En este hito se pide implementar una posible solución a este problema, la cual consiste en ir dando valores al umbral en saltos de 0.3 (este valor ha sido calculado previamente para este grafo). Es decir, $\alpha = 0.3, 0.6, 0.9, \ldots$. A

continuación, para cada posible valor de α se eliminarán las aristas con peso menor al umbral y se calculará $n(\alpha)$, el número de vértices de la componente conexa con mayor número de nodos en el grafo resultante. El valor elegido, α^*, será el menor valor que cumpla $n(\alpha^*) = n(\alpha^* + 0.3)$. Es decir, nos quedaremos con el primer valor de α que no modifique el número de nodos de la mayor componente conexa en la siguiente iteración.

Una vez determinado el valor α^*, consideraremos que cada componente conexa del grafo resultante es un grupo de investigación. Otro aspecto importante a tener en cuenta es que en la normativa de la UCLM se recoge que un grupo de investigación debe tener 5 o más miembros. Por lo tanto, se deben eliminar los grupos con menos de 5 miembros.

Se pide implementar, utilizando como base el hito anterior, un script de MATLAB llamado `Milestone2.m` que realice los siguientes apartados:

a) Determina α^* y $n(\alpha^*)$ para el grafo del Hito 1. A continuación, se deben determinar los grupos de investigación y generar un nuevo grafo por cada grupo de investigación. Guarda en un fichero `.mat` un *cell array* con los grafos generados.

b) Mostrar por pantalla el número de miembros de cada grupo de investigación obtenido usando este procedimiento.

Hito 3: Representación de los grupos de investigación (0.25 puntos)

En este apartado se quiere visualizar los distintos grupos de investigación obtenidos en el hito anterior. Se pide implementar, utilizando como base la plantilla del Hito 1, un script de MATLAB llamado `Milestone3.m` que realice los siguientes apartados:

a) Cargar los grafos de los grupos de investigación almacenados en el cell array generado en el hito anterior.

b) Representar gráficamente (en ventanas separadas) los grupos de investigación obtenidos. Nombra cada ventana como "Research group" seguido del número del grupo $1, 2, 3, \ldots$. Además, establece el grosor de los arcos de manera proporcional al peso de las aristas. Utiliza la formula matemática que consideres adecuada para convertir los pesos de las aristas en el grosor de los arcos en la visualización. La Figura 2 muestra un ejemplo de la representación de uno de estos grupos.

Hito 4: Identificación del IP y de la estructura del grupo de investigación (0.5 puntos)

En este hito se busca estudiar las relaciones internas del grupo de investigación. Para ello, buscaremos obtener un árbol generador que contenga la estructura de dicho grupo. El árbol generador de máximo coste da una representación de las relaciones más intensas entre los miembros del equipo.

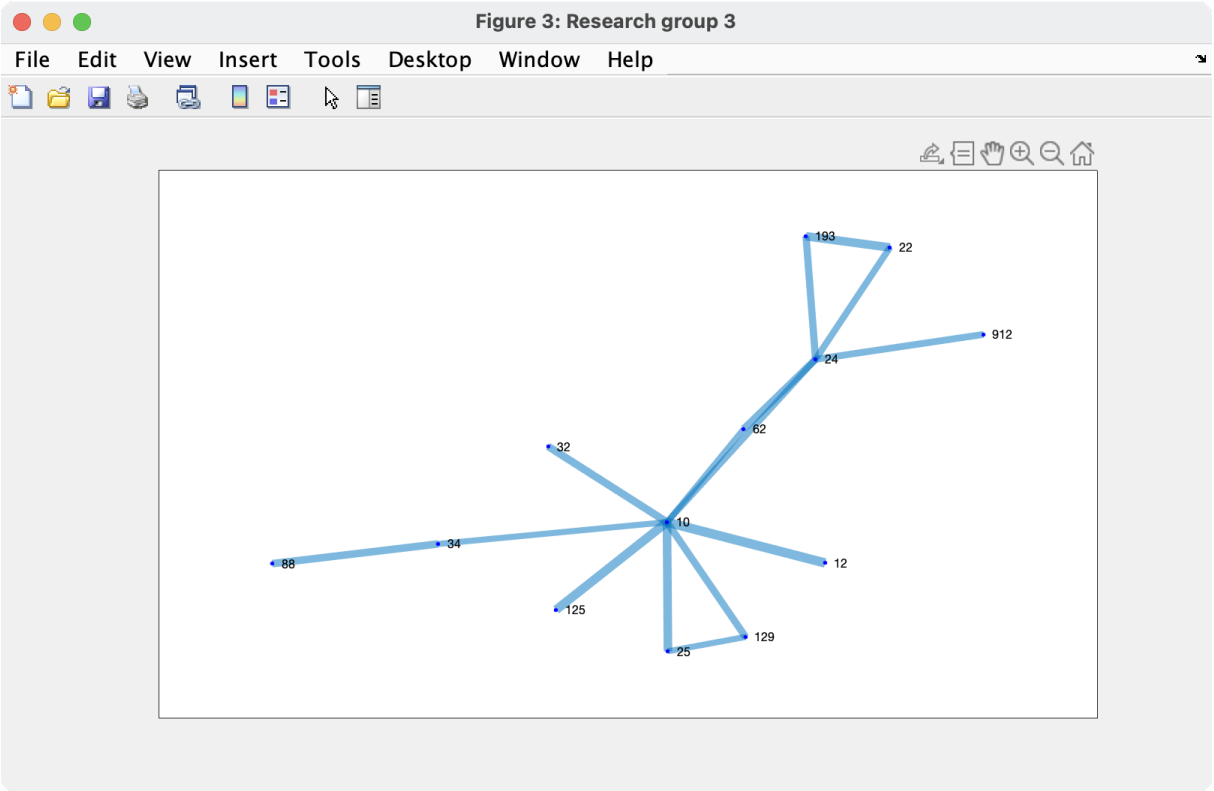

Figura 2: Ejemplo de representación de unos de los grupos obtenidos en el Hito 3

El siguiente paso será determinar que investigadores tienen un mayor peso en este grupo. Se propone definir un índice $\mathcal{P}$ (una etiqueta) que mida la productividad científica de cada investigador en el grupo. Este índice se define como la suma de los pesos de las aristas incidentes con dicho nodo. En un grupo de investigación, la figura del **Investigador Principal (IP)** es el investigador responsable del grupo. Normalmente se corresponde con el investigador con mayor experiencia y, por lo tanto, un mayor índice de productividad. En este hito identificaremos el IP de cada grupo de investigación como aquel miembro del grupo con mayor índice $\mathcal{P}$.

Se pide implementar, utilizando como base la plantilla del Hito 1, un script de MATLAB llamado `Milestone4.m` que cargue los datos necesarios y realice los siguientes apartados:

a) Calcular el índice $\mathcal{P}$ para todos los investigadores de cada grupo de investigación (nodos del grafo).

b) Identificar el IP de cada grupo de investigación.

c) Calcular el árbol generador de **máximo** coste de cada grupo de investigación. Esto consiste en aplicar un algoritmo para obtener el árbol generador de mínimo coste pero considerando el valor negativo de los costes, así en lugar de minimizar logramos maximizar el coste. Utiliza el algoritmo de Prim con raíz en el nodo identificado como IP. No olvides devolver el signo original a los pesos de las aristas en el árbol generador obtenido tras aplicar el algoritmo. Además, añade a la tabla de nodos del árbol generado el índice $\mathcal{P}$ obtenido para cada investigador. Guarda en un fichero `.mat` un *cell array* con los árboles generados.

d) Finalmente, representa el árbol obtenido utilizando una ventana diferente para cada

grupo (como en el hito anterior). En esta representación, marcar los vértices de los IPs detectados anteriormente usando color rojo.

Hito 5: Representación de la estructura del grupo de investigación (0.5 puntos)

En este hito se trata de generar una visualización adecuada para cada grupo de investigación. El IP de un grupo de investigación debe ser la raíz del árbol. Una vez definida la raíz del árbol los distintos miembros se sitúan en diferentes niveles en función de la intensidad de colaboración de estos con el IP y con otros miembros del grupo. De esta forma, se generará una estructura en la que podrán verse distintos subgrupos o temáticas dentro del mismo grupo de investigación.

Se pide implementar, utilizando como base la plantilla del Hito 1, un script de MATLAB llamado `Milestone5.m` que cargue los datos necesarios y realice los siguientes apartados:

a) Representar gráficamente en forma de árbol los diferentes grupos de investigación. La raíz del árbol debe ser el IP del grupo de investigación. Utiliza la función *plot* de MATLAB para ello e investiga sus parámetros. En la representación anterior, etiquetad los vértices con el nombre del investigador.

b) Realiza todas las mejoras que consideres necesarias en la visualización anterior para facilitar la interpretación de los árboles generados. Explica a través de comentarios de texto en el código los cambios realizados. Por ejemplo, se pueden realizar modificaciones para visualizar la importancia de cada miembro de investigación, los distintos subgrupos dentro de un grupo, etc. Se valorará el número de mejoras realizadas en la visualización, así como la adecuación de esta.

Hito 6: Determinar el nombre de los grupos de investigación (0.25 puntos extras)

Finalmente, se propone un último hito opcional que proporcionará 0.25 puntos extra sobre el total de la práctica para aquellos estudiantes que hayan completado toda la práctica y realicen la siguiente mejora. Se deben investigar los distintos grupos de investigación de los IPs obtenidos en el hito anterior y modificar el Hito 5 para que este rotule el nombre de la figuras generadas con el nombre real del grupo de investigación. Utiliza internet para encontrar estos datos. Finalmente, genera un fichero de texto plano llamado `Milestone6.txt` en el que expliques los pasos seguidos para resolver este hito.