Universidad Nacional del Comahue

Facultad de Economía y Administración

Departamento de Matemática

Detección de estructuras latentes en redes complejas mediante técnicas de teoría de grafos y álgebra lineal

Franco Palacios

Directora: Dra. Patricia Janet Caro

Trabajo final para la obtención del título de Licenciado en Matemática

2025

Resumen

El presente trabajo examina la integración de distintos tópicos matemáticos —como la teoría de grafos, el álgebra lineal y el análisis de datos— para el estudio estructural de redes complejas construidas partir de datos relacionales. Se muestran estrategias de detección de comunidades aplicadas tanto a redes simuladas como a redes reales, con el objetivo de mostrar cómo estos enfoques pueden vincularse en un mismo marco analítico. Estas estrategias dependen del objeto del investigador, por lo que los métodos propuestos se aplican según el interés de la determinación de comunidades. Utilizando una base de datos real del sistema de ciencia y tecnología argentino (SICYTAR, 2018) extraída del sitio datos.gob.ar, así como también una red simulada, se aplicaron tres enfoques complementarios: el coloreo de grafos, el agrupamiento espectral y el algoritmo de Girvan–Newman. En primer lugar, se utilizó la técnica del coloreo de grafos como estrategia organizativa para clasificar tareas, atributos y así generar una división preliminar de nodos. Luego, poniendo en práctica conceptos del álgebra lineal como la matriz Laplaciana y sus autovalores, se implementó el agrupamiento espectral para detectar patrones estructurales internos. Por último, se utilizó el algoritmo de Girvan–Newman para identificar comunidades basadas en la eliminación de aristas con un alto valor de intermediación. Todo este proceso siendo hecho con el paquete igraph del software R. Los resultados muestran que la integración de estos métodos permite modelar y optimizar la organización de redes complejas, revelando estructuras internas poco evidentes y ofreciendo una lectura más profunda de las redes y datos relacionales mediante el uso conjunto de herramientas matemáticas. De este modo, este trabajo no solo aporta resultados, sino también una propuesta que puede extenderse a distintos contextos que manejen grandes cantidades de datos, tales como sistemas educativos o instituciones científicas.

Palabras clave: grafos, redes, análisis de datos, matriz Laplaciana, agrupamiento espectral, coloreo, algoritmo de Girvan-Newman.

Abstract

This paper examines the integration of different mathematical topics—such as graph theory, linear algebra, and data analysis—for the structural study of complex networks built from relational data. Community detection strategies are presented, applied to both simulated and real networks, with the goal of demonstrating how these approaches can be linked within a single analytical framework, which is the objective of this thesis. These strategies depend on the researcher's objective, so the proposed methods are applied according to the interest in determining communities. Using a real database of the Argentine Science and Technology System (SICYTAR, 2018) extracted from datos.gob.ar, as well as a simulated network, three complementary approaches were applied: graph coloring, spectral clustering, and the Girvan-Newman algorithm. First, graph coloring was used as an organizational strategy to classify tasks and attributes, thus generating a preliminary node division. Then, applying concepts from linear algebra such as the Laplacian matrix and its eigenvalues, spectral clustering was implemented to detect internal structural patterns. Finally, the Girvan-Newman algorithm was used to identify communities based on the elimination of edges with a high betweenness value. This entire process was performed using the igraph package of the R software. The results show that the integration of these methods allows to model and optimize the organization of complex networks, revealing internal structures that are not obvious and offering a deeper reading of networks and relational data through the joint use of mathematical tools. In this way, this work not only provides results but also a proposal that can be extended to different contexts that handle large amounts of data, such as educational systems or scientific institutions.

keywords: Graphs, Networks, Data Analysis, Laplacian Matrix, Spectral Clustering, Coloring, Girvan-Newman Algorithm.

Índice

Índice de Figuras

1 Introducción

En la actualidad, el análisis de estructuras complejas mediante modelos matemáticos se ha vuelto una herramienta esencial en diversas disciplinas, desde las ciencias sociales y la informática hasta la biología y la economía (Kirman, 2008). Entre los marcos más versátiles y potentes para representar y estudiar relaciones entre entidades se encuentra la teoría de grafos, que permite modelar conexiones, interacciones y flujos de información en sistemas de distinta naturaleza (Satish Kumar et al., 2025). Dentro de este enfoque, uno de los objetivos centrales es la detección de comunidades, entendidas como agrupamientos de nodos más conectados entre sí que con el resto de la red, lo cual permite revelar patrones, jerarquías o funciones latentes. A medida que crecen la cantidad y complejidad de los datos disponibles, se vuelve fundamental contar con métodos que permitan no solo visualizar esas estructuras, sino también descomponerlas, interpretarlas y extraer información significativa (Li et al., 2024). En este trabajo se abordan algunas de las técnicas más representativas para el análisis estructural de redes, poniendo especial énfasis en herramientas provenientes de la teoría de grafos, el álgebra lineal y el análisis de datos.

1.1 Álgebra lineal y grafos

El álgebra lineal y la teoría de grafos son áreas fundamentales de las matemáticas que, aunque inicialmente parecen disjuntas, comparten una relación profunda, especialmente cuando se aplican en el análisis de estructuras complejas. En particular, las matrices, objeto central del álgebra lineal, permiten representar grafos de forma estructurada y eficiente mediante matrices de adyacencia, de incidencia o laplacianas, abriendo así un camino natural entre ambos campos (Bapat, 2010).

Esta representación matricial no solo facilita la manipulación computacional de grafos, sino que también permite aplicar conceptos del álgebra lineal como autovalores, autovectores o transformaciones lineales en el estudio de propiedades estructurales de redes. Chung (1997) señala que las propiedades espectrales de la matriz laplaciana de un grafo contienen información clave sobre su conectividad, particionamiento y dinámica, lo cual ha dado lugar a un área activa conocida como teoría espectral de grafos.

Gracias a este vínculo, herramientas del álgebra lineal se vuelven esenciales en el análisis de grafos sociales, redes neuronales, sistemas distribuidos y otros contextos donde el comportamiento de la estructura depende tanto de sus nodos como de las relaciones entre ellos. Este entrelazamiento teórico permite abordar problemas complejos desde una perspectiva algebraica, facilitando tanto el modelado como la optimización de redes (Newman, 2010).

1.2 Coloreo de grafos

El coloreo de grafos es una estrategia clásica que consiste en asignar colores a los vértices de un grafo de modo que no exista par de vértices adyacentes con el mismo color. Aunque esta técnica surgió inicialmente para problemas como la coloración de mapas, hoy en día sus aplicaciones son mucho más amplias, abarcando ámbitos como la planificación de horarios, la asignación de frecuencias en redes de telecomunicaciones y, de manera creciente, el análisis de datos relacionales (Kannan et al., 2024).

En estos últimos contextos, cuando los datos representan entidades con múltiples relaciones —como usuarios en redes sociales, tareas en sistemas compartidos o recursos que compiten entre sí— modelar dichas relaciones mediante grafos permite aplicar el coloreo para identificar bloques independientes, evitar conflictos y organizar estructuras de datos de forma eficiente. Un ejemplo típico es el uso del coloreo de vértices en compiladores: variables que se usan simultáneamente se conectan mediante aristas, y asignarles "colores" equivale a asignarles registros diferentes (Chaitin, 1982).

Lewis (2016) explica que el coloreo no solo sirve para problemas puramente teóricos, sino que también es crucial para optimización y organización de recursos en sistemas reales. El proceso permite estructurar las relaciones subyacentes y reducir la complejidad de consultas o la gestión de tareas, transformando un conjunto de datos relacionales en un modelo operativo más manejable.

Gracias a esta flexibilidad, el coloreo de grafos se posiciona como una herramienta clave: permite no solo entender relaciones complejas, sino también implementar soluciones prácticas para distribuir tareas, segmentar información y evitar conflictos en ambientes colaborativos.

1.3 Estudio de comunidades en grafos

El análisis de comunidades en grafos consiste en identificar subconjuntos de vértices que están más densamente conectados entre sí que con el resto del grafo. Estos grupos reflejan patrones reales en redes sociales, biológicas, colaborativas o de otro tipo, y permiten revelar estructuras latentes dentro de los datos. En el contexto de datos relacionales, estos nodos pueden representar registros o entidades, mientras que las aristas modelan interacciones, dependencias o similitudes entre ellos (Fortunato, 2010).

Detectar comunidades permite segmentar de manera natural el grafo, facilitando tareas como la comprensión de grupos de usuarios, el descubrimiento de módulos funcionales en redes biológicas o el agrupamiento de tareas relacionadas en sistemas distribuidos. Entre los métodos más conocidos se encuentran el algoritmo de Girvan–Newman, que utiliza la eliminación iterativa de aristas con alta intermediación, y los enfoques basados en la optimización de modularidad, que miden la calidad de una partición según la densidad de enlaces internos frente a una distribución aleatoria. También existen métodos algebraicos como el agrupamiento espectral, que se basa en el análisis de los autovalores y autovectores de matrices asociadas al grafo —como la laplaciana—, permitiendo detectar comunidades a través de propiedades estructurales profundas del grafo (Newman, 2010).

Fortunato (2010) también sostiene que las comunidades representan "compartimentos relativamente independientes dentro de un sistema", y que su identificación es clave para entender la función y organización de redes complejas. Además, describe y compara distintos enfoques, desde algoritmos jerárquicos hasta métodos espectrales y de optimización, destacando sus fortalezas y limitaciones en términos de precisión y escalabilidad.

Así, el estudio de comunidades no solo aporta un diagnóstico estructural del grafo, sino que también ofrece un mecanismo para resolver problemas operativos en datos relacionales. Al identificar subconjuntos densamente interconectados, se facilita la construcción de soluciones modulares, el diseño de consultas eficientes o la agrupación de registros relacionados (Aggarwal & Wang, 2010).

1.4 Objetivo general y objetivos específicos

Como se ha expuesto en las secciones anteriores, la teoría de grafos ofrece un marco versátil para modelar sistemas complejos mediante nodos y relaciones. Su vínculo con el álgebra lineal permite representar de forma matemática la estructura de redes y aplicar herramientas analíticas potentes, como las derivadas del análisis espectral. En este contexto, el estudio del coloreo de grafos surge como una forma de organización estructural, útil en tareas de asignación o planificación, mientras que la detección de comunidades se orienta a revelar la segmentación interna de redes, identificando agrupamientos naturales o funcionales entre nodos.

Partiendo de estas ideas, el objetivo general de esta tesis es explorar y aplicar distintas estrategias de análisis estructural (detección de comunidades) en grafos, tanto sobre redes simuladas como sobre datos relacionales reales, con énfasis en tres enfoques: el coloreo de grafos como paso organizativo inicial, el algoritmo de Girvan–Newman como método clásico de detección de comunidades, y el agrupamiento espectral, complementado con k-means, como alternativa basada en herramientas algebraicas.

Aunque el coloreo no se clasifica formalmente como técnica de detección de comunidades, se lo incorpora como un primer paso que permite estructurar el grafo y observar cómo una partición basada en restricciones locales puede influir en el análisis posterior. A través de esta combinación de enfoques, se busca no solo mostrar el funcionamiento técnico de cada método, sino también reflexionar sobre su aplicabilidad en contextos donde los datos representan relaciones reales, como es el caso del conjunto utilizado en esta investigación. Dicho esto, los objetivos específicos son los siguientes:
1. Organizar las redes para la identificación de estructuras y facilitar la interpretación visual de las mismas utilizando coloreo de grafos.
2. Detectar comunidades y patrones latentes con el método de agrupamiento espectral, como herramienta algebraica basada en propiedades de matrices asociadas a los grafos.
3. Identificar comunidades con el algoritmo de Girvan-Newman, como método clásico fundamentado en la eliminación iterativa de aristas con alta intermediación.

En síntesis, este trabajo pretende integrar herramientas provenientes de distintas áreas —teoría de grafos, álgebra lineal y análisis de datos— en un mismo marco analítico, mostrando cómo pueden colaborar para ofrecer una lectura más rica y estructurada de las redes complejas.

1.5 Estructura del trabajo de tesis

El desarrollo de esta tesis estará dividido en cinco secciones principales. En primer lugar, la introducción, donde fueron planteados los temas y objetivos de la tesis. Luego, el marco teórico presenta los conceptos fundamentales que sustentan el trabajo, agrupados en cuatro apartados: teoría de grafos, coloreo de grafos, representaciones matriciales y estudio de comunidades. A continuación, la sección de metodología detalla los procedimientos aplicados, incluyendo la implementación de algoritmos, el uso de software específico, las características de la base de datos utilizada y los pasos seguidos para el análisis. La sección de resultados expone los grafos obtenidos, las comunidades detectadas y observaciones relevantes sobre su estructura. Finalmente, en la conclusión, se realiza un cierre del trabajo con reflexiones generales, limitaciones y posibles líneas futuras de investigación.

2 Marco Teórico

2.1 Conceptos de grafos

Un grafo es una terna $G = (V, U, \Phi)$ que consiste en dos conjuntos no vacíos y disjuntos, V y U , de elementos llamados vértices y aristas respectivamente, y de una función Φ, frecuentemente llamada relación de adyacencia, que asocia a cada arista de U un par no ordenado de vértices (no necesariamente distintos) de G. Se puede extender la definición de grafo para cuando $U = \varnothing$, en este caso la terna asociada es $(V, \varnothing, \varnothing)$ y se tiene el grafo discreto (Braicovich et al., 2009).

Si u es una arista del grafo G, a y b son vértices tales que $\Phi(u) = (a, b)$, entonces se dice que la arista u tiene extremos en los vértices a y b . Una arista en la que coinciden ambos extremos es llamada bucle o lazo. Los vértices que son extremos de una misma arista, que no es bucle, se llaman vértices adyacentes. Se aceptan aristas diferentes pero con los mismos extremos, las mismas son llamadas aristas múltiples. Cuando un grafo no tenga este tipo de aristas, diremos que es un grafo simple.

Sea $G = (V, U, \Phi)$ un grafo, diremos que G es:

- Finito si los conjuntos V y U son conjuntos finitos.
- De orden n si es finito y el número de elementos de V es igual a n.
- Trivial si es discreto y de orden 1.

El grado de un vértice v del grafo G , se nota $gr(v)$ y es el número de aristas con extremos en v, se cuenta doble cada bucle. Un vértice se dice aislado cuando ninguna arista lo tiene por extremo, por lo tanto el grado de dicho vértice es igual a cero. Un vértice se dice pendiente cuando su grado es igual a uno. Un grafo simple sin bucles es completo si cada vértice es adyacente a todos los restantes, notaremos K_n al grafo completo de orden n.

Observación 1: Esta definición formal considera en su nomenclatura la relación de adyacencia anteriormente definida. Es importante indicar que, por comodidad, en lo que sigue y cuando no haya lugar a confusión, se utilizará una notación menos formal para nombrar a los grafos, la que se indica a continuación (Braicovich et al., 2009).

Notaremos como $G = (V, U)$ al grafo de vértices en el conjunto V y aristas en el conjunto U, es decir mediante un par y no una terna. Obviamente en el conjunto U las aristas estarán indicadas mediante pares no ordenados, para así conocer cuáles son los extremos de cada una de las mismas. Además, a partir de este punto, se emplearán como sinónimos los términos redes y grafos, nodos y vértices, así como enlaces y aristas.

En los estudios de redes sociales, a menudo es importante saber si es posible alcanzar algún nodo n_i desde otro nodo n_j. Si es posible, también puede ser interesante saber cuántas maneras hay de hacerlo, y cuál de esos modos es óptimo con respecto a uno de varios criterios (Iacobucci, 2013). Por ejemplo, podríamos desear entender la comunicación de información entre empleados en una organización. Una consideración importante es si la información originada en un empleado podría eventualmente alcanzar a todos los otros empleados, y si es así, cuántas aristas debe atravesar para llegar allí. También se podría considerar si hay múltiples rutas que un mensaje podría tomar para ir de un empleado a otro, y si algunos de esos caminos son más o menos "eficientes"(Iacobucci, 2013).

Un camino (en inglés, walk) es una secuencia de nodos y aristas, comenzando y terminando con nodos, en la cual cada nodo es incidente con las aristas que lo siguen y lo preceden en la secuencia. El listado de un camino, denotado por W, es una secuencia alternada de nodos y aristas incidentes comenzando y terminando con nodos. El nodo de inicio y el nodo del final pueden ser diferentes. Además, algunos nodos pueden estar incluidos más de una vez, y algunas aristas pueden estar incluidas más de una vez. La longitud de un camino es el número de ocurrencias de aristas en él. Si una arista está incluida más de una vez en el camino, se cuenta cada vez que aparece. Formalizado, tenemos:

Definición 2.1.1. *Dado un grafo $G = (V, U)$ un camino es una secuencia de vértices $v_1v_2...v_{n+1}$ y aristas $a_1a_2...a_n$ tales que $a_i = \{v_i, v_{i+1}\}$ para todo $i \in \{1, ..., n\}$. La longitud del camino es n. (Iacobucci, 2013).*

Debido a que los grafos (simples) tienen como máximo una arista entre cada par de nodos, no hay ambigüedad sobre qué arista está entre cualquier par de nodos, y un camino puede describirse simplemente listando los nodos involucrados y excluyendo las aristas. El nodo inicial y el nodo final de un camino son el primer y el último nodo de W y son referidos como el principio y el final de W. El inverso de un camino, denotado por W^{-1}, es el camino W exactamente con el orden opuesto, usando los mismos nodos y aristas.

Un camino es el tipo de secuencia más general de nodos adyacentes, ya que no hay restricción sobre que nodos y aristas puedan estar incluidos (aparte de la proximidad de los nodos). Tipos especiales de caminos, que consideramos a continuación, son más restrictivos en que los nodos o aristas no sean usados más de una vez.

Cuando el nodo inicial y el final coinciden se dice que el camino es cerrado, y si además no se repiten vértices ni aristas, se lo llama ciclo. Los recorridos y los caminos simples son caminos con características especiales. Un recorrido (en inglés, trail) es un camino en el cual todas las aristas son distintas, aunque algunos nodos pueden estar incluidos más de una vez. En el ejemplo de comunicaciones, un recorrido significa que ningún lazo de comunicación se utiliza más de una vez. La longitud de un recorrido es el número de aristas en él.

Un camino simple (en inglés, path) es un camino en el cual todos los nodos y todas las aristas son distintos. Por ejemplo, un camino simple a través de una red de comunicación significa que ningún actor es informado más de una vez. La longitud de un camino simple es el número de aristas en él.

Nótese que todo camino simple es un recorrido, y todo recorrido es un camino. Así que cualquier par de nodos conectado por un camino simple también está conectado por un recorrido y por un camino. Así, un camino es lo más general y un camino simple es lo menos general dentro de los tipos de "ruta" en un grafo. Dado que todos los caminos simples son caminos (pero sin repetir nodos ni aristas), un camino simple probablemente sea más corto comparado con un camino o un recorrido. En un camino simple en la red de comunicaciones, ningún empleado es informado más de una vez, y ningún par de empleados discute el asunto más de una vez. En aplicaciones a redes sociales, a menudo nos enfocaremos en caminos simples en lugar de caminos (Šumak & Pušnik, 2023).

Una propiedad muy importante de un par de nodos es si hay o no un camino simple entre ellos. Si hay un camino simple entre los nodos n_i y n_j, entonces se dice que n_i y n_j son alcanzables. Por ejemplo, si consideramos una red de comunicaciones entre personas en la que las aristas en un grafo representan canales para la transmisión de mensajes entre personas, entonces si dos actores son alcanzables, es posible que un mensaje viaje de un actor al otro pasando el mensaje a través de intermediarios. Si dos actores no son alcanzables, entonces no hay un camino simple entre ellos, y no hay forma de que el mensaje viaje de un actor a otro (Holme, 2005).

Ejemplo 2.1.2. Sea $G = (V, U)$, el conjunto $V = \{v_1, v_2, v_3, v_4\}$ y el conjunto
$U = \{(v_1, v_2), (v_1, v_2), (v_3, v_2), (v_2, v_2), (v_1, v_4), (v_3, v_4)\}$.

Con estos datos el grafo se representa a continuación:

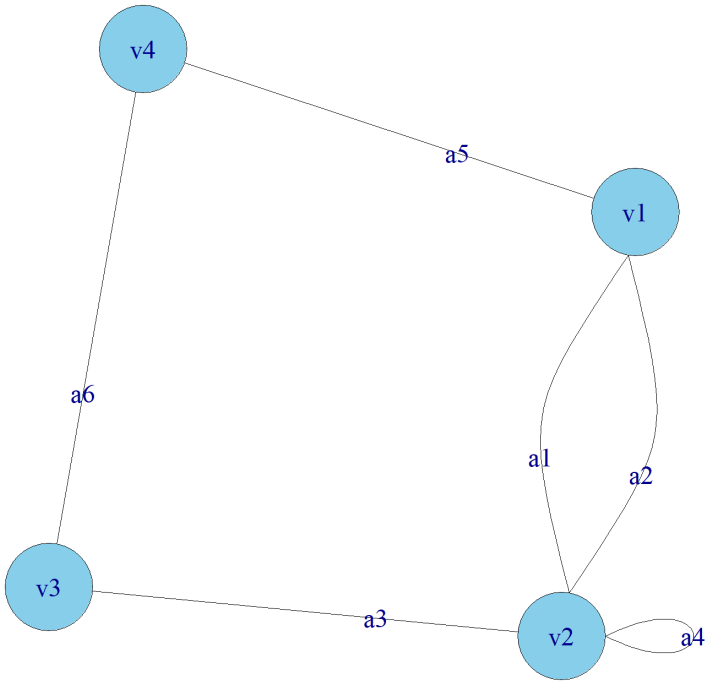

Figura 1: Grafo G

Es sumamente importante destacar que no interesa la ubicación de los vértices ni la forma o longitud de las aristas en los grafos, pues la información que podemos sacar de los mismos no está relacionada con dichos parámetros.

Se puede observar en este grafo que:

- La arista a_4 es un bucle, las aristas a_1 y a_2 tienen los mismos extremos, por lo que son aristas múltiples, también denominadas aristas paralelas. Por este motivo, es decir por tener aristas múltiples, no es simple el grafo G.
- Los vértices v_1 y v_2, por ejemplo, son adyacentes y en cambio, los vértices v_1 y v_3 no lo son.
- No existen en este grafo vértices aislados ni tampoco pendientes, ya que los valores de los grados son: $gr(v_1) = 3$, $gr(v_2) = 5$, $gr(v_3) = 2$ y $gr(v_4) = 2$
- El camino C: v_2, a_3, v_3, a_6, v_4 es un camino simple al no repetirse aristas ni vértices.

2.1.1 Grafo planar

Una representación en el plano de un grafo $G = (V, U)$ es una función f tal que:

- A cada vértice $v \in V$ le hace corresponder un punto del plano $\mathbf{R}^2$.
- A cada arista $a \in U$ le hace corresponder una curva simple con extremos en los puntos del plano correspondientes a los puntos extremos de a y tal que la curva no contiene otros puntos correspondientes a vértices del grafo.

Un grafo G admite distintas representaciones, sin embargo es importante destacar que una representación determina un único grafo.

Definición 2.1.3. *Un grafo G se dice grafo planar si G admite una representación en el plano tal que curvas correspondientes a aristas distintas no se cortan salvo, tal vez, en sus puntos extremos. Una representación tal se dice una representación plana de G o una inmersión en el plano de G. (Braicovich et al., 2009).*

Definición 2.1.4. *Un grafo plano es un grafo planar con una dada representación plana. Se puede pensar que un grafo plano es una representación plana particular. (Braicovich et al., 2009).*

Al grafo que cumple con la condición de planaridad lo llamaremos indistintamente grafo plano o grafo planar.

Definición 2.1.5. *Un grafo $G = (V, U)$ es bipartito si el conjunto V puede ser particionado en dos subconjuntos, V_1 y V_2 , tal que cada arista de G tiene un extremo en el conjunto V_1 y otro extremo en el conjunto V_2. En particular, si G es un grafo tal que cada vértice del conjunto V_2 es adyacente a cada vértice del conjunto V_2, entonces el grafo G es bipartito completo. Notaremos con $K_{r,s}$, al grafo bipartito completo, donde $|V_1| = r$ y $|V_2| = s$. (Braicovich et al., 2009)*

Estas dos nociones se generalizan para el caso en que el conjunto V puede ser particionado en k conjuntos $V_1, V_2, ..., V_k$. En este caso, se dice que G es k-partito y k-partito completo respectivamente. En este último caso, si se tiene que $|V_i| = n_i$, $1 \leq i \leq k$, el grafo será notado $K_{n_1,n_2,...,n_k}$. Este tipo de grafos tiene distintas aplicaciones, una de ellas referida a la asignación, puede ser de tareas-empleados, de máquinas-operarios,etc.

Ejemplo 2.1.6. En una empresa deben ser realizadas 7 tareas $(t_1, t_2, t_3, ..., t_7)$ y se dispone de 5 operarios $(o_1, o_2, ..., o_5)$. Si el operario o_1 es capaz de realizar las tareas t_1 y t_3 , el operario o_2 es capaz de realizar las tareas t_4 y t_6, o_3 la tarea t_2 , o_4 las tareas t_5 y t_6 y finalmente el operario o_5 es capaz de realizar las tareas t_4 y t_7. Esta situación puede ser representada mediante el siguiente grafo bipartito.

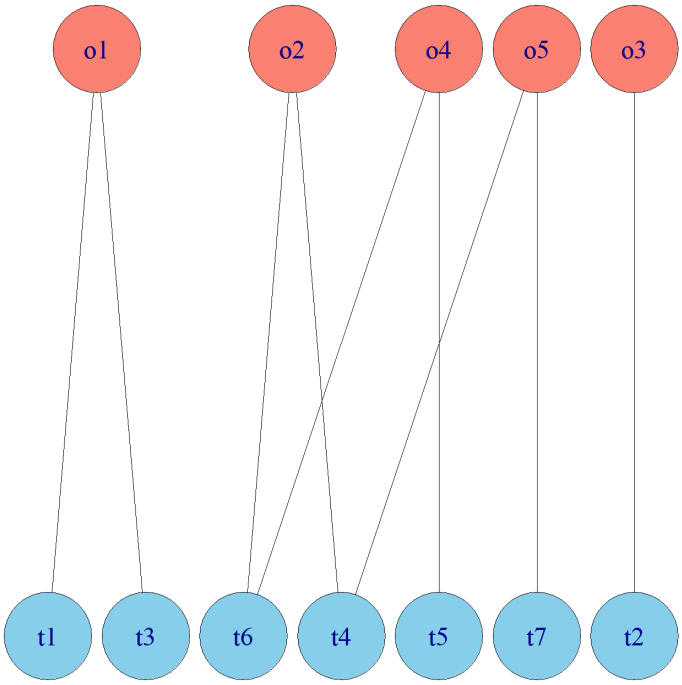

Figura 2: Grafo bipartito

En el caso que cada uno de los 5 operarios esté en condiciones de realizar las 7 tareas, el grafo bipartito sería completo, notándolo en este caso $K_{5,7}$, o equivalentemente $K_{7,5}$.

2.1.2 Subgrafos

Dado un grafo $G = (V, U)$ se dice que el grafo $G' = (V', U')$ es un subgrafo de G si se tiene que: $V' \subseteq V$ y $U' \subseteq U$. Dado un grafo G , podemos decir que G' es subgrafo:

- cubriente si $V' = V$.
- inducido por U' si está constituido por las aristas de U' y los vértices de G sobre los cuales inciden estas aristas.
- inducido por V' si está constituido por los vértices de V' y las aristas de G cuyos extremos pertenecen a V'.

Sea $G = (V, U)$ un grafo, $V' \subseteq V$ y $U' \subseteq U$. Notaremos con $G - V'$ al subgrafo de G inducido por $V - V'$ y con $G - U'$ al subgrafo obtenido a partir de G eliminando las aristas pertenecientes a U'. Particularizando este caso, tomando conjuntos V' y U' de cardinalidad igual a 1, tenemos:

- El subgrafo restante respecto de un vértice v, es el subgrafo obtenido a partir del grafo original omitiendo el vértice v y todas las aristas incidentes en él, simbolizaremos a este subgrafo como $\tilde{G}_v$.

- De manera similar al caso anterior, se dice que el grafo que se obtiene al quitar una arista a al grafo G es llamado subgrafo restante respecto de una arista a y se nota como $\tilde{G}_a$.

Ejemplo 2.1.7. Sea el grafo $G = (V, U)$ del ejemplo 2.1.2, el que se grafica a continuación:

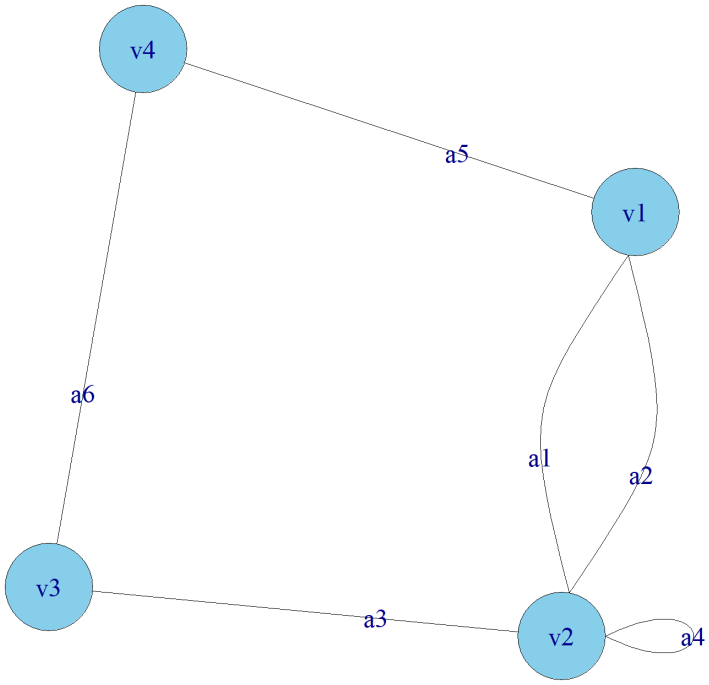

Figura 3: Grafo del ejemplo 2.1.2

Pueden considerarse, entre otros, los siguientes subgrafos:

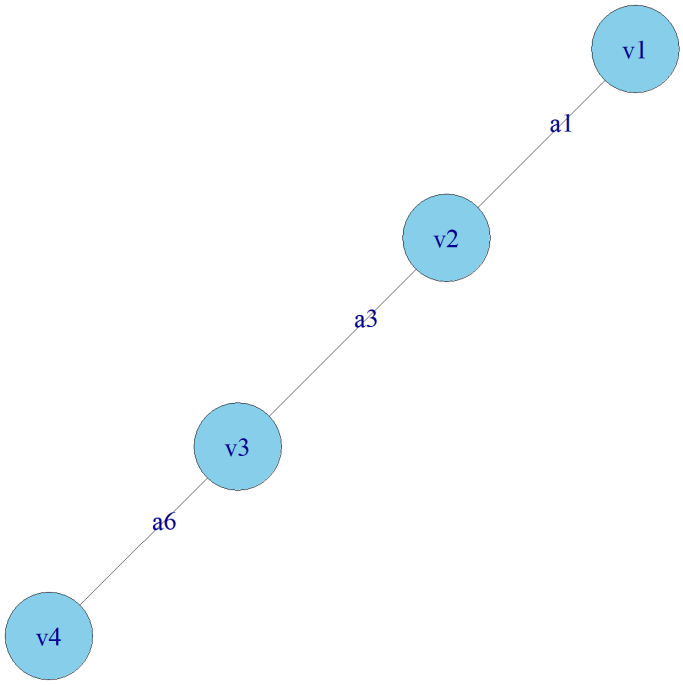

Figura 4: Subgrafo cubriente inducido por U'

Este subgrafo es cubriente pues el conjunto de vértices del mismo coincide con el conjunto V, y se dice inducido por el conjunto $U' = \{a_1, a_3, a_6\}$. Otro subgrafo que puede considerarse es el siguiente:

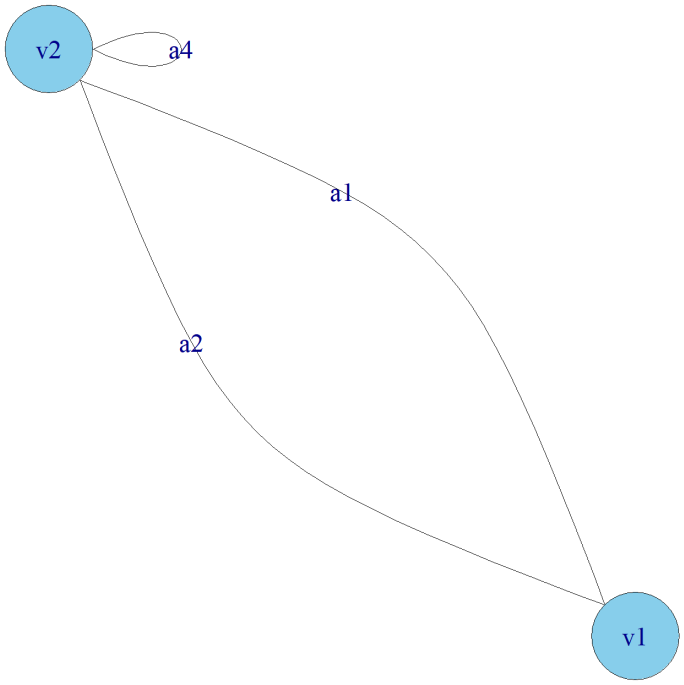

Figura 5: Subgrafo no cubriente inducido por V'

Este subgrafo no es cubriente pues el conjunto de vértices del mismo no coincide con el conjunto V. En este caso el subgrafo es inducido por el conjunto $V' = \{v_1, v_2\}$ y puede notarse como $G - \{v_3, v_4\}$.

Definición 2.1.8. *Un grafo G es conexo si es trivial o, equivalentemente, para cada par de vértices de G existe al menos un [camino] que los une. Caso contrario, G es disconexo o no conexo.*

Ejemplo 2.1.9. Damos a continuación un grafo conexo y uno no conexo:

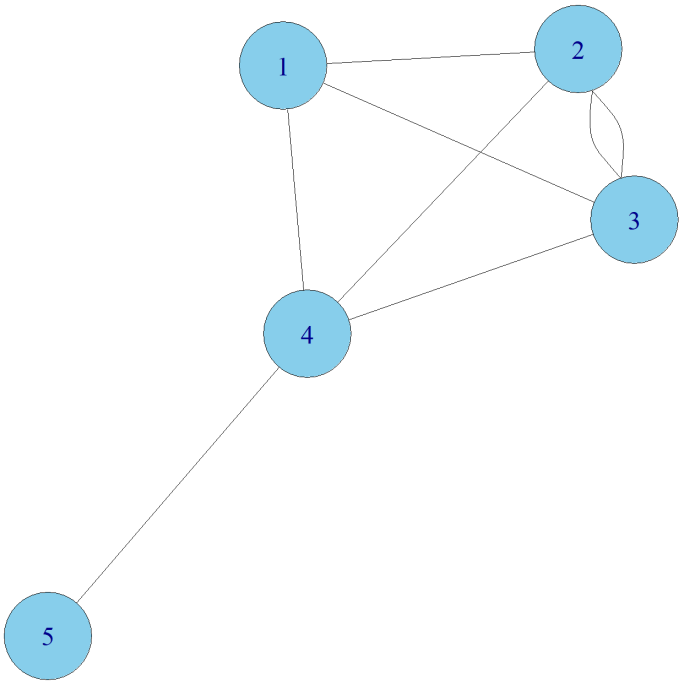

Figura 6: Grafo del ejemplo 2.1.9

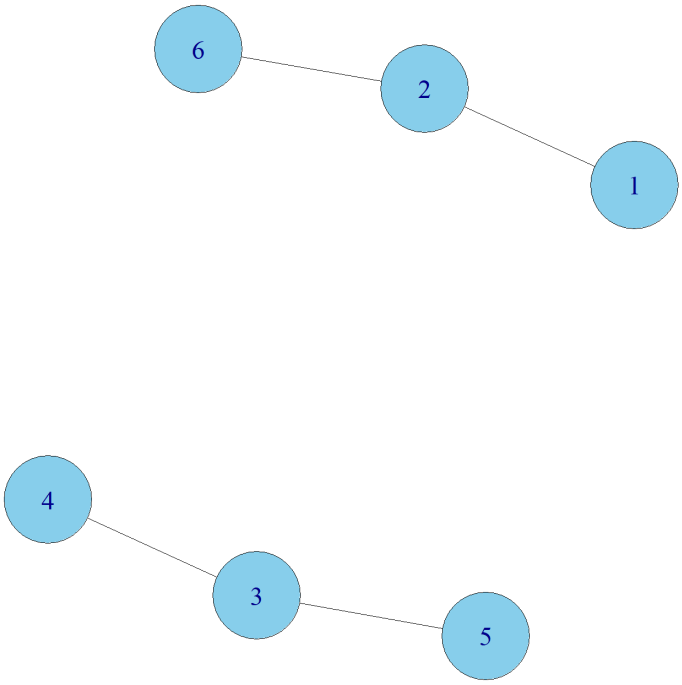

Figura 7: Grafo formado por 2 componentes conexas

Un grafo disconexo consiste en dos o más subgrafos conexos, cada uno de estos subgrafos es una componente conexa.

Si G_1, G_2,.....,G_t son las t componentes conexas de G, se tiene que: $G = \bigcup_{i=1}^{t} G_i$. El vértice v del grafo conexo G es un istmo si el subgrafo $\tilde{G}_v$ no es conexo. La arista a del grafo conexo G es un puente si el subgrafo $\tilde{G}_a$ no es conexo.

2.1.4 Grafo de similitud

Dado un conjunto de puntos de datos $x_1, ..., x_n$ y alguna noción de similitud $s_{ij} \geq 0$ entre todos los pares de puntos de datos x_i y x_j, el objetivo intuitivo del agrupamiento (clustering) es dividir los puntos de datos en varios grupos tales que los puntos dentro del mismo grupo sean similares entre sí y los puntos en diferentes grupos sean disímiles entre sí. Si no tenemos más información que las similitudes entre los puntos de datos, una buena manera de representar los datos es en forma de un grafo de similitud $G = (V, E)$. Cada vértice v_i en este grafo representa un punto de datos x_i. Dos vértices están conectados si la similitud s_{ij} entre los puntos de datos correspondientes x_i y x_j es positiva o mayor que un cierto umbral, y la arista se pondera con s_{ij} (von Luxburg, 2007).

Ejemplo 2.1.10. Supongamos que tenemos 4 puntos de datos x_1, x_2, x_3, x_4 y las similitudes entre pares de puntos están definidas con la siguiente matriz $S = (s_{ij})$ (La diagonal es 0 porque lógicamente

no se mide similitud de un punto consigo mismo).

$$S = \begin{pmatrix} 0 & 0.8 & 0.2 & 0 \\ 0.8 & 0 & 0.4 & 0.1 \\ 0.2 & 0.4 & 0 & 0.7 \\ 0 & 0.1 & 0.7 & 0 \end{pmatrix}$$

Si tomamos como umbral el valor 0.2 esto significa que si $s_{ij} \geq 0.2$ hay arista y si $s_{ij} < 0.2$ no la hay. Ahora, los vértices serán $V = \{v_1, v_2, v_3, v_4\}$ donde cada uno representa un punto x_i y las aristas son $U = \{(v_1, v_2), (v_1, v_3), (v_2, v_3), (v_3, v_4)\}$. Luego nos queda el grafo de similitud.

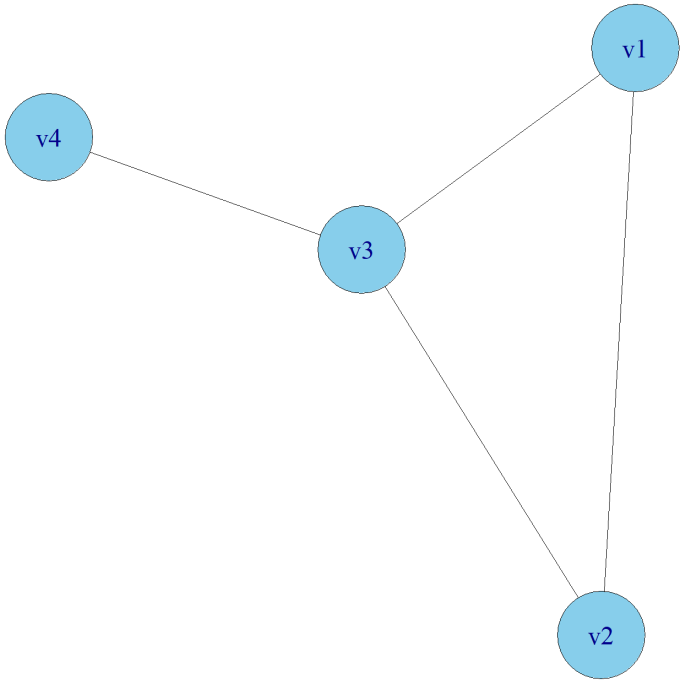

Figura 8: Grafo de similitud

2.1.3 Grafos dirigidos

Antes de seguir con el coloreo, se repasará de manera breve el concepto de grafos dirigidos.
Dado $G = (V, U, \Phi)$ en el cual los elementos del conjunto U no son aristas sino que son arcos, es decir pares ordenados, diremos que G es un grafo dirigido y lo llamaremos digrafo. En este caso la función Φ es llamada relación de incidencia (Braicovich et al., 2009).

Si u es un arco, a y b son vértices, no necesariamente distintos, tales que $\Phi(u) = [a, b]$, entonces se dice que u tiene extremo inicial en a y extremo final en b. Si un arco tiene igual extremo inicial que final es denominado bucle o lazo.

Las nociones de digrafo finito, de orden n, discreto y trivial son análogas al caso no dirigido.

Al igual que en el caso de grafos, utilizaremos una notación menos formal cuando no haya lugar a confusión. Notaremos a los digrafos como $G = (V, U)$, donde V es el conjunto de vértices y U es el conjunto de arcos, los que se indican mediante pares ordenados de vértices.

2.2 Coloreo

El coloreo de grafos es una técnica matemática y computacional que ha encontrado aplicaciones en diversas áreas. Este método se basa en la famosa conjetura de los cuatro colores, que establece que cualquier mapa puede ser coloreado con solo cuatro colores de manera que no haya dos regiones adyacentes con el mismo color (Appel & Haken, 1977).

Un tema relevante donde se puede apreciar la versatilidad del coloreo de grafos en distintas aplicaciones es el estudio sobre la coloración de grafos y su uso en geografía. Este articulo (Carrasco-Pilco, Burgos-Cevallos, Jurado-Liberona & Nymoen-Bonilla, 2021) investiga la coloración de grafos y establece que cualquier mapa puede ser coloreado con solo cuatro colores justamente como se dijo antes, utilizando cuatro métodos conocidos en la demostración del Teorema de los Cuatro Colores.

Un problema que parece haber sido mencionado por el matemático alemán August Moebius, en 1840, y ser consecuencia directa de una hipótesis de los fabricantes de mapas es lo que dio origen a la conjetura de los cuatro colores, la misma, de manera formal dice lo siguiente: "Supuesto que cada país está constituido por una única región conexa y que toda frontera entre países esté formada por arcos de curva (no las hay constituidas por un solo punto) todo mapa sobre un plano, o equivalentemente sobre la superficie de una esfera, puede colorearse utilizando a lo sumo cuatro colores y de forma que países limítrofes tengan colores distintos".

Pocos meses después de terminar sus estudios en el University College of London, Francis Guthrie escribió a su hermano Frederick, quién estaba todavía en el "College" y era discípulo del matemático Augustus De Morgan. En su carta, Francis hacía notar a Frederick que bastaban cuatro colores para colorear tales mapas, preguntándole si sería posible demostrar esto matemáticamente. Frederick no lo sabía y se lo consultó a su profesor De Morgan, quién lo ignoraba también. El primer testimonio escrito sobre esta conjetura es la carta que data del año 1852 donde Augustus De Morgan le pregunta al matemático Hamilton sobre esta cuestión. En 1860 Charles Peirce anunció haberla probado pero su manuscrito nunca fue publicado.

El asunto no fue de interés general hasta el año 1878, cuando Arthur Cayley comunica en una reunión de la London Mathematical Society que había sido incapaz de resolverlo y entonces sí la conjetura se constituyó en uno de los más famosos desafíos matemáticos de la época.

Antes de pasar un año, Alfred Kempe, abogado y miembro de la Sociedad, publicó un artículo en el que afirmaba haber demostrado que la conjetura era verdadera. Pero once años más tarde, el matemático británico Persy John Heawood encontró un error y mostró que era válido si se sustituía cuatro por cinco colores. A pesar de haber sido incompleto el razonamiento de Kempe contenía la mayor parte de las ideas fundamentales que un siglo después habrían de llevar a la demostración correcta.

En el año 1969 los matemáticos Ore y Stemple demostraron su validez para todos los mapas con a lo sumo cuarenta países. Finalmente pudo ser demostrada en términos generales, en el año 1976, por Kenneth Appel y Wolfgang Haken, quienes recurrieron al uso del computador. Esta demostración, según sus propios autores, sugiere que existe un límite para lo conseguible por métodos puramente teóricos y agregan que con anterioridad se ha subestimado la necesidad de métodos computacionales en las demostraciones matemáticas. Mencionan tener la esperanza de que su trabajo contribuya a avanzar en esta dirección y que esta ampliación de las técnicas válidas de demostración justifique el gran esfuerzo dedicado durante tantos años a la demostración del teorema de los cuatro colores.

Todos los esfuerzos realizados en la comunidad matemática con el fin de decidir respecto de la validez de esta conjetura impulsaron el desarrollo de la topología combinatoria y llevaron al estudio de los grafos planares. Se ejemplifica la cuestión que ha sido mencionada tomando un mapa no coloreado de América del Sur (Braicovich et al., 2009), el mismo se presenta a continuación:

Figura 9: Mapa de América del Sur

Si se intenta colorear este mapa con 3 colores de manera que países limítrofes tengan asignado distinto color, se comprueba que es imposible hacerlo. Se puede ver, por ejemplo, que Argentina, Bolivia, Paraguay y Brasil limitan todos ellos entre sí por lo que se necesitan como mínimo 4 colores para poder colorear el mapa.

Se puede comprobar, tal cual se muestra en esta segunda imagen del mapa, que es posible colorear con 4 colores distintos de manera que países limítrofes tengan diferente color.

Figura 10: Mapa de América del Sur a color

Esta coloración no es única, resulta sencillo observar, por ejemplo, que Ecuador puede ser coloreado con el mismo color que Venezuela, Paraguay y Guayana Francesa. También se puede cambiar el color asignado a Chile, dándole el mismo color que a Paraguay.

Es posible verificar que todo mapa en el que los límites sean segmentos y no puntos, trazado sobre una hoja de papel puede ser representado mediante un grafo planar, donde los vértices indican los países y las aristas unen a aquellos que son limítrofes. A partir de esto, resulta que la ya confirmada conjetura es equivalente a la siguiente proposición: "Para colorear los vértices de un grafo planar es suficiente utilizar cuatro colores". Así, el grafo que se presenta a continuación e indica los límites entre los países de América del Sur, nos muestra la vinculación que existe entre esta conjetura y la coloración de un grafo planar:

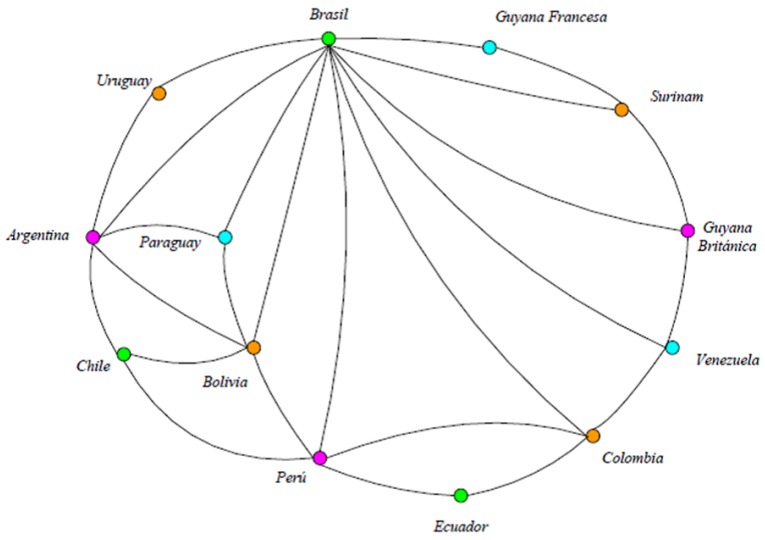

Figura 11: Grafo planar

El coloreo de un grafo consiste en darle colores a los vértices de manera tal que a vértices adyacentes correspondan colores diferentes.

La cantidad mínima de colores necesarios para colorear los vértices de un grafo G es llamada número cromático de G y se la nota como $\chi(G)$.

Los vértices del grafo G pueden ser agrupados en subconjuntos disjuntos de acuerdo con el color que le fuera dado, estos subconjuntos son conjuntos independientes de vértices. La cantidad de conjuntos independientes $C_1, C_2, \ldots, C_k$ en que puede ser particionado el conjunto de vértices de un grafo G, se llama k-coloración de G y el número cromático de G será el menor número natural k tal que G admita una k-coloración.

Ejemplo 2.2.1. Si tomamos el grafo anterior podemos ver que la partición del conjunto de vértices en cuatro conjuntos independientes de acuerdo con el color asignado es la siguiente:

$C_1 = \{$Brasil, Chile, Ecuador$\}$
$C_2 = \{$Uruguay, Bolivia, Colombia, Surinam$\}$
$C_3 = \{$Argentina, Perú, Guyana Británica$\}$
$C_4 = \{$Paraguay, Venezuela, Guyana Francesa$\}$

En este caso y como ya se aclaró en el punto anterior, esta partición no es la única que se puede realizar, se presenta una distinta a continuación:

$C_1 = \{$Brasil, Chile$\}$
$C_2 = \{$Uruguay, Paraguay, Perú, Guyana Británica$\}$
$C_3 = \{$Argentina, Colombia, Guyana Francesa$\}$
$C_4 = \{$Bolivia, Ecuador, Venezuela, Surinam$\}$

Teorema 2.2.2. *Si el valor p es el grado máximo entre los vértices del grafo G, entonces se tiene que: $\chi(G) \leq p + 1$ (Braicovich et al., 2009).*

Dem: Es trivial si el número de vértices del grafo G es igual a 1 ó 2. Para los otros casos, supondremos válido el enunciado para todos los grafos con $(n-1)$ vértices y demostraremos que también es válido para aquellos grafos con n vértices.

Sea G un grafo con n vértices, si quitamos un vértice v cualquiera de G, el grafo $\tilde{G}_v$ resultante tiene $(n-1)$ vértices y por hipótesis inductiva, admite una $(p-1)$-coloración. Sean $C_1, C_2, ..., C_p, C_{p+1}$ los conjuntos partición de $V(G)$.

Como $g(v) \leq p$, hay uno de estos conjuntos C_i $(1 \leq i \leq p+1)$ que no contiene vértices adyacentes a v, entonces podemos agregar este vértice a dicho conjunto C_i para poder, entonces, obtener una $(p-1)$-coloración para el grafo G, como queríamos demostrar.

Corolario 2.2.3. *Si p es el grado máximo entre los vértices de un grafo completo de orden $p+1$, entonces $\chi(K_{p+1}) = p+1$ (Braicovich et al., 2009).*

En un coloreo de vértices, cada color induce un conjunto independiente y como el grafo es completo, todo par de nodos es adyacente, así que cualquier conjunto independiente tiene a lo sumo un vértice, por lo tanto, cada color podrá asignarse como máximo a uno de ellos, lo que hace directo comprobar que $\chi(K_{p+1}) = p+1$.

Ejemplo 2.2.4. Sea el grafo completo de orden 4, ya coloreado:

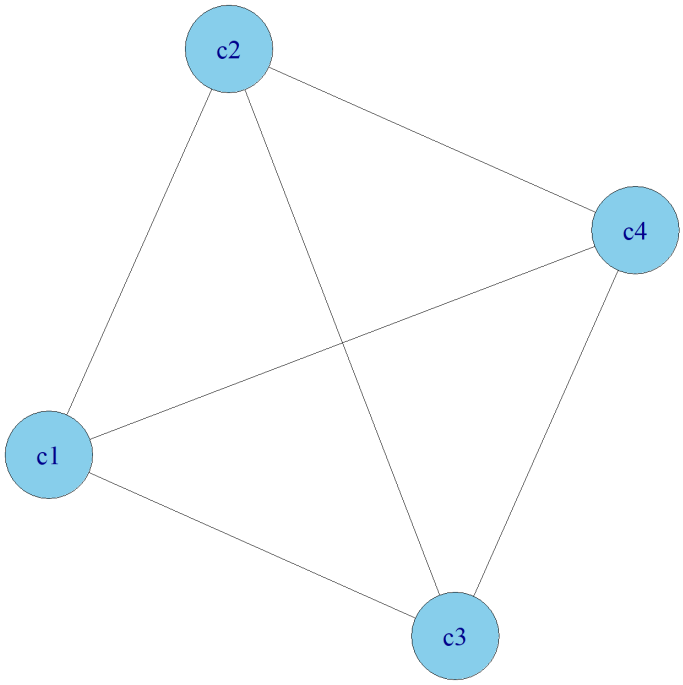

Figura 12: Grafo del ejemplo 2.2.4

Es directo comprobar que es necesario asignar colores distintos a cada uno de los vértices, pues en un grafo completo todos sus vértices se encuentran conectados entre sí, lo que lleva a que cada uno de los conjuntos partición sea unitario. En este caso particular, $\chi(K_4) = 4$.

Teorema 2.2.5. *(Teorema de König)*
Sea un grafo G, se tiene: $\chi(G) \leq 2$, si y sólo si, G no contiene ciclos de longitud impar (Braicovich et al., 2009).

Dem ida:

En este sentido es directa la demostración ya que todo ciclo de longitud impar necesita 3 colores y $\chi(S) \leq \chi(G)$ si el grafo S es un subgrafo de G. Por lo tanto, se puede afirmar que si G contiene algún ciclo impar, $\chi(G) \geq 3$, en este caso se tiene por hipótesis que $\chi(G) \leq 2$, lo que asegura que el grafo G no contiene ciclos de longitud impar.

Dem vuelta:

Recíprocamente, se sabe por hipótesis que G no contiene ciclos de longitud impar y sin pérdida de generalidad se supone que G es conexo, ya que en caso contrario, se aplicaría el proceso a cada una de las componentes conexas.

Sea $B_0 = v_0$, donde v_0 es un vértice cualquiera de G. Llámese B_1 al conjunto de vértices adyacentes a v_0, B_2 al conjunto de vértices distintos de v_0 y adyacentes a los vértices de B_1 y general, definimos a B_i como el conjunto de vértices adyacentes a los vértices de B_{i-1} que no están contenidos en B_j, donde $j < i$.

Así, se define una 2-coloración del grafo G, como sigue:

$$C_1 = B_0 \cup B_2 \cup ... \text{ y } C_2 = B_1 \cup B_3 \cup ...$$

Para verificar que C_1 es independiente basta observar que cualquier vértice de C_1 está unido a v_0 por una cadena de longitud par. Si dos vértices de C_1 fuesen adyacentes completarían con la arista que los une un ciclo de longitud impar, lo que es, por hipótesis, imposible.

Un razonamiento análogo, prueba la independencia de C_2, quedando así demostrado el teorema.

Corolario 2.2.6. *El número cromático de cualquier grafo bipartito completo de orden n, $K_{n,n}$, es igual a 2, es decir $\chi(K_{n,n}) = 2$ (Braicovich et al., 2009).*

Corolario 2.2.7. *El número cromático de cualquier grafo bipartito $G_{n,m}$ es igual a 2, es decir $\chi(G_{n,m}) = 2$ (Braicovich et al., 2009).*

Es directa la demostración de los dos últimos corolarios, ya que por ser grafos bipartitos no contienen ciclos de longitud impar. Se ejemplifican a continuación los corolarios anteriormente presentados.

Ejemplo 2.2.8. Se presenta el grafo $G = K_{3,3}$, ya coloreado, siendo su número cromático igual a 2.

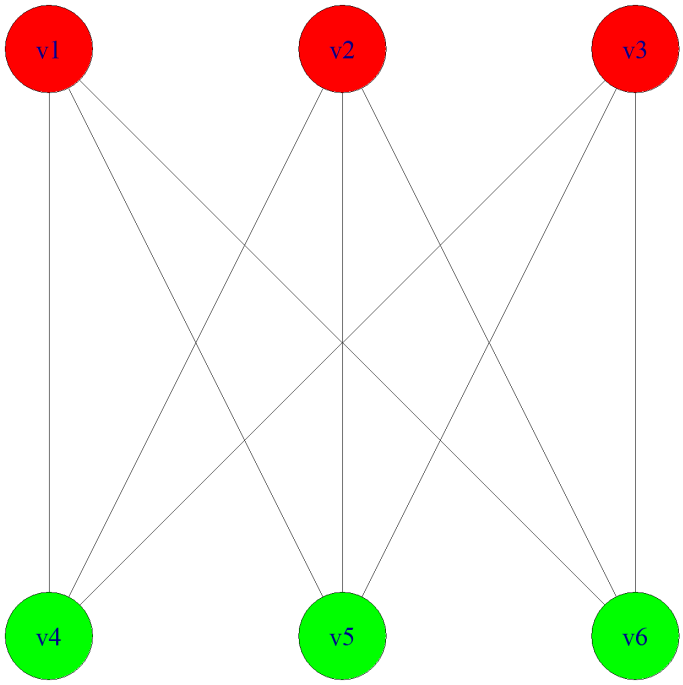

Figura 13: Grafo del ejemplo 2.2.8

Ejemplo 2.2.9. En el siguiente grafo bipartito se puede ver claramente que sus vértices quedan particionados en 2 clases determinadas por la coloración.

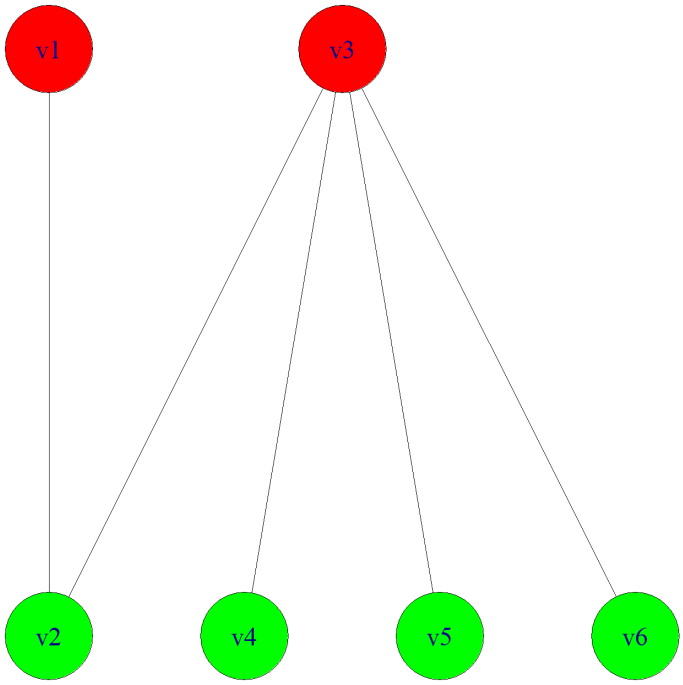

Figura 14: Grafo del ejemplo 2.2.9

Teorema 2.2.10. *Es posible colorear con dos colores las regiones del plano determinadas por un número finito de rectas incluidas en él, de manera que regiones adyacentes tengan colores diferentes (Braicovich et al., 2009).*

Dem:

Realizaremos esta demostración utilizando inducción matemática sobre el número de rectas incluidas en un plano α.

El resultado es trivial si el número de rectas es $n = 1$, se colorea cada semiplano determinado por la recta de un color distinto, quedando entonces coloreado el plano con 2 colores.

Resta suponer verdadero el enunciado para $(n - 1)$ rectas y demostrar que es también válido para n rectas.

El suponer válido para $(n - 1)$ rectas, significa que ya se encuentran bicoloreadas las regiones del plano determinadas por ellas. Al trazar

una recta más en dicho plano, resulta que en uno de los semiplanos determinados por esta n-ésima recta se debe conservar la coloración previa y en el semiplano opuesto se deben invertir los colores, quedando así probado el teorema.

A continuación se presenta un plano con distintas regiones determinadas por rectas, las que ya han sido coloreadas de manera que regiones adyacentes tengan distinto color.

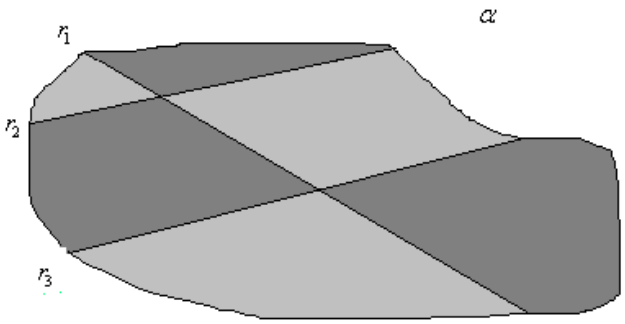

Figura 15: Plano con regiones adyacentes de distinto color

2.3 Coloreo de grafos en asignación de tareas

Se presenta a continuación un caso ilustrativo, en el que se observa otro tipo de aplicación del coloreo de grafos.

La aplicación del algoritmo de coloreo de grafos puede optimizar la asignación de tareas en ambientes colaborativos. Se compararon dos escenarios: uno con una distribución aleatoria de tareas y otro con una distribución organizada tras la aplicación del algoritmo de coloreo. La muestra utilizada fue el departamento de Matemática de la Universidad del Comahue (UNCO), con 67 docentes y tareas divididas en cuatro grupos: docencia, investigación, extensión y gestión. Utilizando el software R, se generaron grafos bipartitos para representar las asignaciones de tareas, y se analizaron métricas de centralidad del autovector y de centralidad de grado. Los resultados mostraron que el algoritmo de coloreo logra una distribución más igual y balanceada de las tareas, evitando la sobrecarga en algunos individuos y asegurando que todos los docentes tengan tareas asignadas de manera justa.

El caso estudiado aquí se centra en aplicar estos conceptos a ámbitos donde la asignación de tareas entre personas sea recurrente, demostrando cómo los principios del coloreo de grafos pueden ser utilizados para mejorar la gestión y distribución de tareas en ambientes donde un grupo de individuos colaboran en pos de un objetivo común o, dicho de otra manera, en un ambiente colaborativo. Por lo tanto, el objetivo en este análisis es demostrar cómo la aplicación del algoritmo de coloreo de grafos puede optimizar la organización y el equilibrio de las tareas en un ambiente colaborativo. En términos prácticos, esto significa asegurar que todas las tareas estén asignadas de manera igualitaria, evitando la sobrecarga de trabajo en individuos específicos y garantizando que no haya nadie sin tareas asignadas. Este enfoque busca mejorar la eficiencia y la claridad en la gestión de tareas, creando un ambiente de trabajo más ordenado y balanceado.

Aquí se introducen las definiciones de centralidad de grado y de autovector que seran usadas a continuación.

Centralidad de grado: Es igual al número de vínculos o enlaces que tiene un nodo con otros nodos en el grafo de la red. La ecuación que representa el número de vecinos más cercanos para grafos no dirigidos está dada por la siguiente expresión, donde A es la matriz de adyacencia (véase Definición 2.4.8):

$$C_D(i) = k_{(i)} = \sum_j A_{ij} = \sum_j A_{ji}$$

Es decir, $C_D(i)$ es el grado del vértice i (Perez, 2022).

Centralidad del Autovector: Se encuentra asociado al mayor autovalor de la matriz de adyacencia y mide la influencia que tiene cada nodo en una red. La centralidad del autovector definido de esta manera otorga a cada nodo una centralidad que dependetanto de la cantidad como de la calidad de

sus conexiones. La importancia de un nodo depende de la importancia de sus vecinos.

$$v_i \longleftarrow \sum_j A_{ij} v_j$$

donde A es la matriz de adjacencia.

$$v_i = \frac{1}{\lambda} \sum_j A_{ij} v_j (Av = \lambda v)$$

Se selecciona el autovector asociado al mayor autovalor $\lambda = \lambda_1$, $v = v_1$

El presente caso analizado busca demostrar cómo la aplicación del algoritmo de coloreo de grafos puede optimizar la asignación de tareas en ambientes colaborativos. Para ello, se plantearon dos escenarios: el primero con una asignación de tareas aleatoria, donde algunos individuos están sobrecargados de deberes y otros no tienen ninguno; y el segundo, donde se aplica el algoritmo de coloreo para distribuir las tareas de manera más ordenada e igualitaria. La muestra utilizada fue el departamento de Matemática de la Universidad del Comahue (UNCO), que cuenta con 67 docentes. Las tareas se dividen en cuatro grupos: docencia, investigación, extensión y gestión. Cada grupo posee cuatro tareas, enumeradas de la siguiente manera:
Docencia (D)
D1. Dictar clases
D2. Tomar examen
D3. Dar consulta
D4. Armar programa de materia

Investigación (I)
I1. Publicar artículos científicos
I2. Desarrollar proyectos
I3. Revisión por pares
I4. Tratamiento de software y herramientas

Extensión (E)
E1. Diseñar proyectos comunitarios
E2. Organizar capacitaciones
E3. Divulgación científica
E4. Ofrecer asesorías

Gestión (G)
G1. Supervisor de tesis
G2. Dirección de carrera
G3. Coordinación del departamento
G4. Planificación académica

Usando el software R y el paquete writexl, se crearon dos archivos Excel que sirvieron como base de datos para cada escenario. En el primer escenario, las dieciséis tareas se asignaron de manera aleatoria a los 67 docentes. Como era de esperarse, algunos docentes quedaron sobrecargados con varias tareas, incluso pertenecientes al mismo grupo, mientras que otros no tenían más de una tarea asignada. En el segundo escenario, la base de datos se ajustó para asegurar que cada docente tuviera cuatro tareas asignadas, una de cada grupo. Posteriormente, se utilizaron los paquetes igraph y visNetwork en R para elaborar grafos bipartitos. Los nodos representaban a los docentes y las tareas, con los grafos denominados "Grafo 1" y "Grafo 2", correspondientes al primer y segundo escenario, respectivamente. Cada grupo tuvo un color asignado y así los nodos de las tareas se colorearon según el grupo al que pertenecían: De esta manera, cada docente tenía una tarea de cada grupo, garantizando que los nodos

vecinos de las tareas fueran de diferente color, aplicando así el algoritmo de coloreo. Finalmente, utilizando el paquete ggplot2 en R, se generaron gráficos para analizar las diferencias entre ambos escenarios, Se realizaron gráficos de barras mostrando la distribución de tareas por docente, tablas de contingencia que indicaban la asignación de tareas por grupo y diagramas de caja (box plots) que presentaban la centralidad del autovector entre los grafos. Se eligió esta métrica debido a que las de cercanía e intermediación están más relacionadas con los caminos de un grafo, lo que no tenía sentido en el contexto de grafos bipartitos.

En la Figura 12 se muestra el Grafo 1, correspondiente a la situación antes de la asignación de tareas, mientras que la Figura 13 presenta un acercamiento al nodo del docente 64, donde se observa que solo tiene asignada una tarea. Por su parte, la Figura 14 muestra el Grafo 2, correspondiente a la situación después de la asignación de tareas, y la Figura 15 amplía nuevamente el nodo del docente 64, evidenciando que ahora cuenta con una tarea de cada grupo.

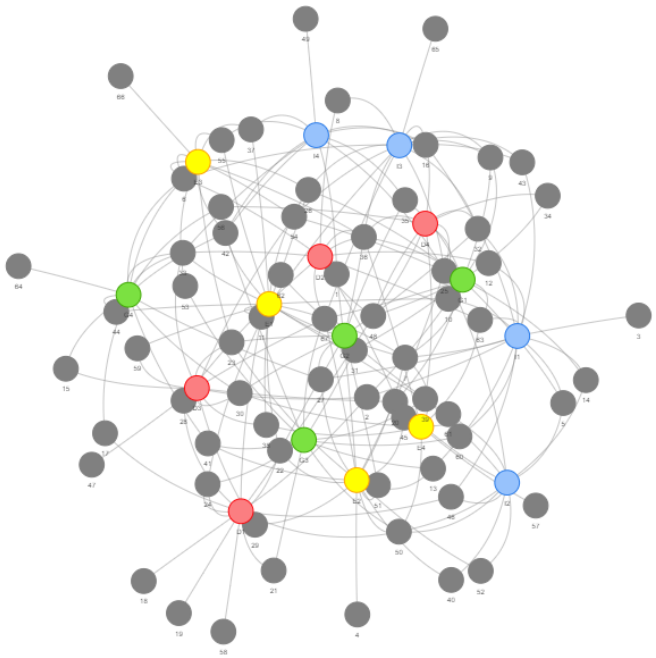

Figura 16: Grafo 1

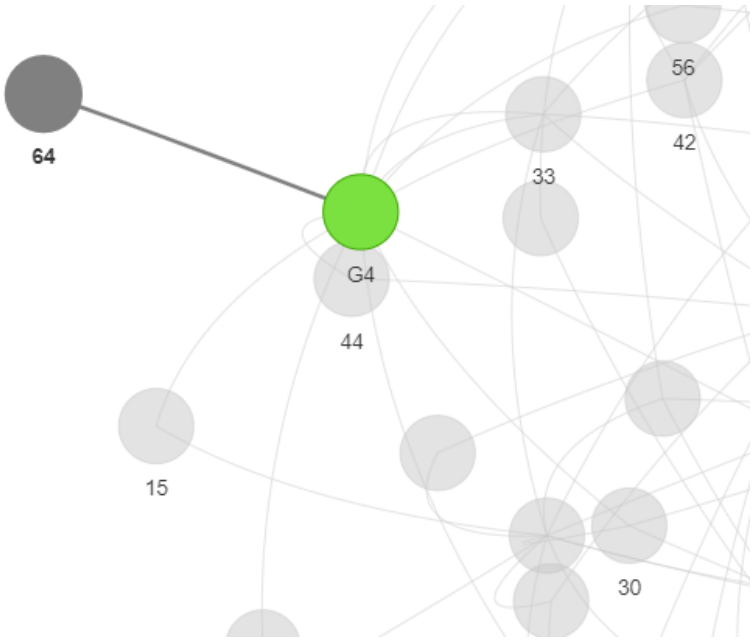

Figura 17: Grafo 1 ampliado

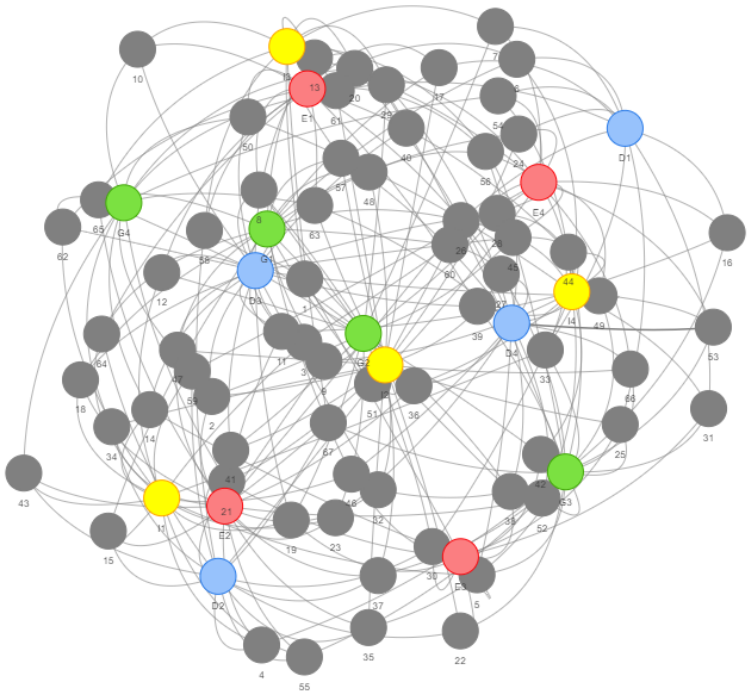

Figura 18: Grafo 2

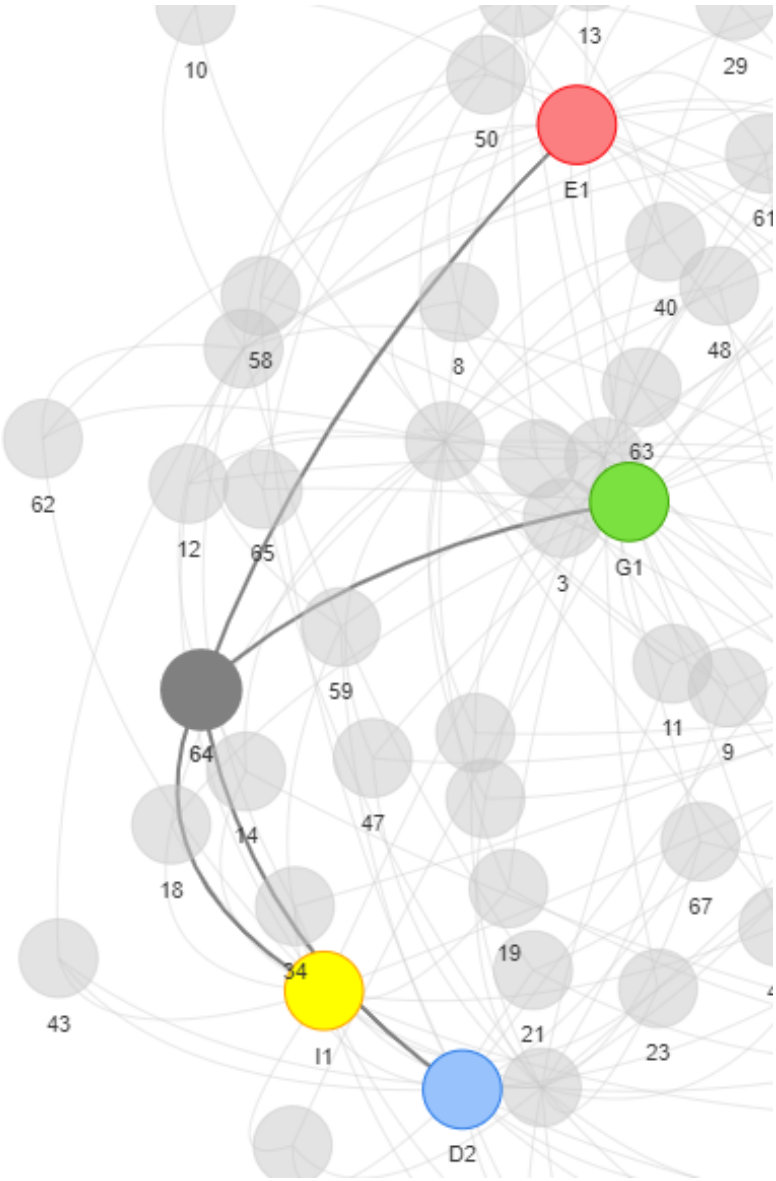

Figura 19: Grafo 2 ampliado

A continuación, veremos una serie de gráficos que nos permiten comparar y analizar ambos grafos

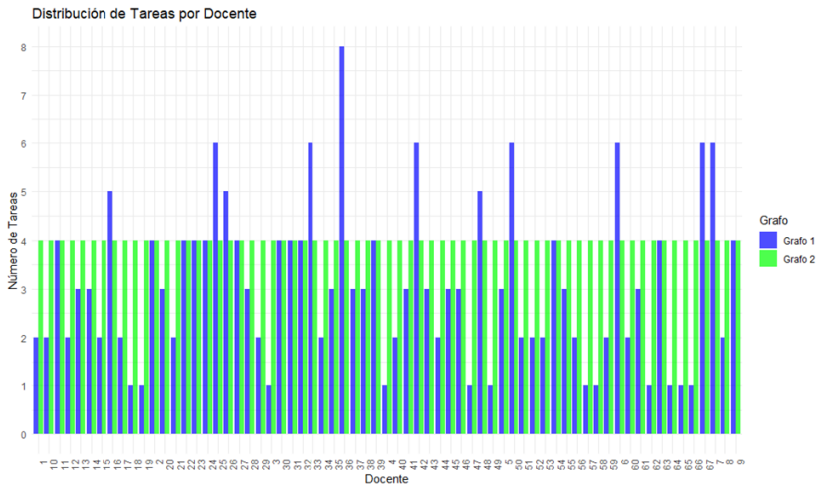

Figura 20: Gráfico de barras

Este gráfico de barras indica la centralidad de grado de cada nodo de los docentes, es decir ilustra la cantidad de tareas asignadas a cada uno de ellos, comparando los dos escenarios diferentes: el Grafo 1, que representa la asignación aleatoria y desordenada de tareas, y el Grafo 2, que muestra la distribución ordenada tras la aplicación del algoritmo de coloreo. En el eje vertical se presenta el número de Tareas, mientras que en el eje horizontal se enumeran los docentes. Las barras azules corresponden al Grafo 1 y las barras verdes al Grafo 2. En el Grafo 1, se observa una distribución irregular, con algunos docentes cargados con múltiples tareas, incluso del mismo grupo, mientras que otros tienen solo una tarea asignada. Esta desorganización genera un desequilibrio significativo en la carga de trabajo. En contraste, el Grafo 2 muestra una distribución mucho más igualitaria y uniforme de las tareas entre los docentes. Esto nos sugiere que la aplicación del algoritmo de coloreo logra una asignación más balanceada, evitando la sobrecarga en algunos individuos y asegurando que todos los docentes tengan tareas asignadas de la misma manera. El gráfico refleja la eficacia del algoritmo para optimizar la distribución de tareas en un ambiente colaborativo, promoviendo así un entorno de trabajo más equilibrado y ordenado. El siguiente box plot compara la distribución de la centralidad del autovector entre el Grafo 1 y el Grafo 2. En el eje y se presenta la centralidad del autovector, mientras que el eje x diferencia entre los dos grafos.

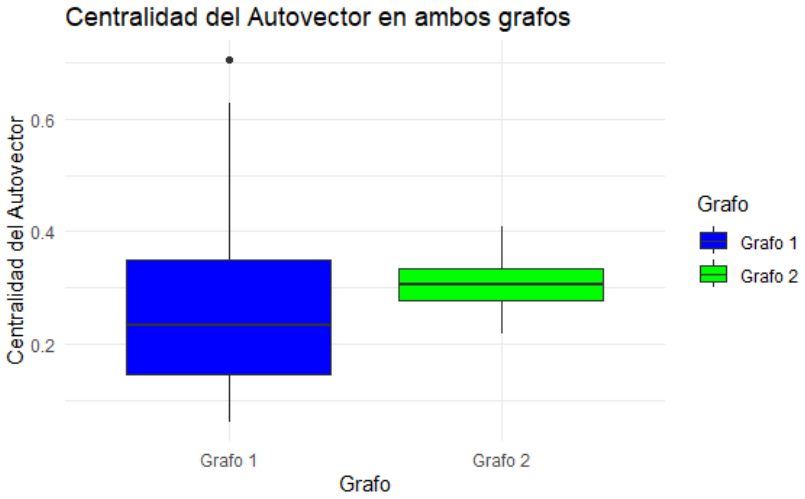

Figura 21: Boxplot de la centralidad del autovector

El box plot del Grafo 1 (coloreado en azul), muestra una mayor variabilidad en los valores de centralidad del autovector, con algunos valores atípicos por encima de 0.6. La mediana es más baja en comparación con el Grafo 2. Esto nos sugiere que, en el escenario desordenado, hay una dispersión más amplia en la importancia de los nodos, con algunos docentes teniendo un rol más central y otros menos influyentes. En contraste, el box plot del Grafo 2 (coloreado en verde), muestra una distribución más concentrada de los valores de centralidad del autovector. La mediana es más alta, indicando que, en promedio, los docentes tienen una mayor centralidad en el grafo ordenado, y la variabilidad es menor. Esto nos indica que la aplicación del algoritmo de coloreo no solo distribuye las tareas de manera más igualitaria, sino que también equilibra la importancia de los nodos dentro del grafo. Por lo tanto, este gráfico nos reafirma la efectividad del algoritmo de coloreo en la asignación de tareas, logrando una distribución más equilibrada y ordenada de las responsabilidades entre los docentes, lo que se refleja en la centralidad del autovector más estable del Grafo 2. En definitiva, tras mirar los gráficos, se puede concluir que la aplicación del algoritmo de coloreo equilibra la asignación de tareas en términos absolutos, Por último, también se puede apreciar con la siguiente tabla cuantos docentes había asignados con tareas de cada grupo (podían repetirse entre grupos) y como esto se ordenó tras aplicar el algoritmo de coloreo, ya que ahora se tiene a todos los docentes con tareas asignadas de todos los grupos.

	D	E	I	G
Grafo 1	48	54	51	52
Grafo 2	67	67	67	67

Figura 22: Tabla de docentes con sus tareas

Finalizando, podemos decir que, en este marco, se ha demostrado cómo la aplicación del algoritmo de coloreo de grafos puede optimizar la asignación de tareas en ambientes colaborativos. A través de la comparación de dos escenarios, uno con una distribución aleatoria y otro con una distribución organizada post-algoritmo, se evidenció que la implementación del coloreo de grafos logra una distribución más igualitaria y balanceada de las tareas entre los docentes del departamento de Matemática de la Universidad del Comahue (UNCO). Los resultados mostraron que, al aplicar el algoritmo, no solo se evita la sobrecarga de trabajo en algunos individuos, sino que también se asegura que todos los

docentes tengan tareas asignadas de manera justa y ordenada. Además, al analizar las métricas de centralidad del autovector se confirmó que el algoritmo de coloreo equilibra la importancia de los nodos dentro del grafo. En conclusión, el uso del algoritmo de coloreo de grafos se presenta como una herramienta efectiva para mejorar la organización y la eficiencia en la asignación de tareas en ambientes colaborativos, promoviendo así un entorno de trabajo más justo y productivo.

2.4 Matrices

En esta sección se desarrollará teoría de matrices haciendo especial foco en la matriz laplaciana, la cual es clave para el agrupamiento espectral, pero antes de ello se repasarán algunas definiciones previas a tener en cuenta.

Definición 2.4.1. *Una matriz $A \in \mathbb{R}^{n\times n}$ se dice simétrica si $A_{ij} = A_{ji}\ \forall\ 1 \leq i,\ j \leq n$ o, equivalentemente, si $A = A^t$. (Jerónimo, Sabia & Tesauri, 2008)*

Definición 2.4.2. *Una matriz simétrica $A \in \mathbb{R}^{n\times n}$ es definida positiva si $x^T Ax > 0$ para todo $x \in \mathbb{R}^n$ distinto de cero. Es semidefinida positiva si $x^T Ax \geq 0$ para todo $x \in \mathbb{R}^n$ distinto de cero. Es indefinida si existen vectores $y, z \in \mathbb{R}^n$ tal que $y^T Ay < 0 < z^T Az$.(Horn & Johnson, 2013).*

La idea de la perpendicularidad para los vectores en $\mathbb{R}^n$, se le llama ortogonalidad. En $\mathbb{R}^2$ o $\mathbb{R}^3$, dos vectores u y v distintos de cero son perpendiculares si el ángulo θ entre ellos es un ángulo recto; esto es, si $\theta = \pi/2$ radianes o $90°$. Por tanto, $\frac{u\cdot v}{\|u\|\|v\|} = \cos 90° = 0$, y se concluye que $u \cdot v = 0$. Esto motiva la siguiente definición (Poole, 2010).

Definición 2.4.3. *Dos vectores u y v en $\mathbb{R}^n$ son mutuamente ortogonales si $u \cdot v = 0$*

Definición 2.4.4. *Un conjunto de vectores $\{v_1, v_2, ..., v_k\}$en $\mathbb{R}^n$ se llama conjunto ortogonal si todos los pares de vectores distintos en el conjunto son ortogonales; esto es, si*

$$v_i \cdot v_j = 0 \text{ siempre que } i \neq j \text{ para } i, j = 1, 2, ..., k$$

Definición 2.4.5. *Un conjunto de vectores en $\mathbb{R}^n$ es un conjunto ortonormal si es un conjunto ortogonal de vectores unitarios. Una base ortonormal para un subespacio W de $\mathbb{R}^n$ es una base de W que es un conjunto ortonormal.*

Definición 2.4.6. *Una matriz Q de $n \times n$ cuyas columnas forman un conjunto ortonormal se llama matriz ortogonal.*

Definición 2.4.7. *Una matriz cuadrada A es diagonalizable ortogonalmente si existe una matriz ortogonal Q y una matriz diagonal D tales que $Q^T AQ = D$.*

2.4.1 Matriz de adyacencia

Definición 2.4.8. *Sea G un grafo de orden n, la matriz adyacencia de G se define como $A(G) = (a_{ij})$ donde a_{ij} es el número de aristas de la forma (i, j), eventualmente $i = j$.*

A continuación se presenta un grafo de 5 vértices, por lo que la matriz adyacencia será de orden 5. Se presenta a continuación la misma, a modo de ejemplo.

Ejemplo 2.4.9. Dado el grafo G siguiente

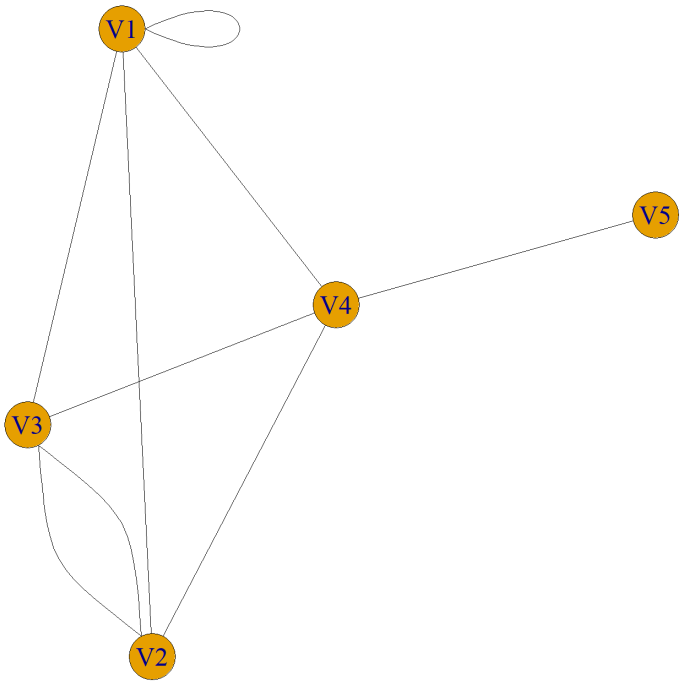

Figura 23: Grafo G del ejemplo 2.4.9

La matriz adyacencia del grafo es la siguiente:

$$A(G) = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 2 & 1 & 0 \\ 1 & 2 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

De la definición de matriz adyacencia, resulta que $A(G)$ es simétrica, ya que las aristas se indican mediante pares no ordenados.

2.4.2 Matriz de incidencia

Antes de abordar la matriz laplaciana, que es el eje central de este apartado, se presentan las definiciones y ejemplos de la matriz de incidencia, siguiendo a Braicovich (2009).

Definición 2.4.10. *Sea G un grafo sin bucles, de vértices x_i y de aristas u_j, $1 \leq i \leq n$, $1 \leq j \leq m$. La matriz de incidencia de G es $\underline{B}(G) = (\underline{b}_{ij})$, donde:*

$$\underline{b}_{ij} = \left\{ \begin{array}{cc} 1 & si\ u_j\ incide\ en\ x_i \\ 0 & caso\ contrario \end{array} \right\}$$

Ejemplo 2.4.11. Sea el siguiente grafo G

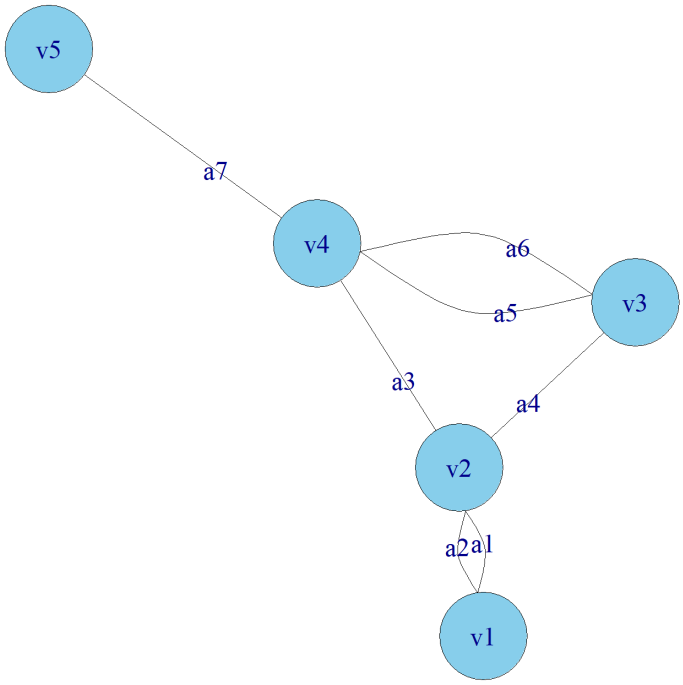

Figura 24: Grafo del ejemplo 2.4.11

La matriz de incidencia del mismo es de orden 5×7, ya que cada una de las filas representa un vértice y cada una de las columnas representa una arista:

$$\underline{\mathrm{B}}(G) = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Definición 2.4.12. *Sea el grafo dirigido G sin bucles, de vértices x_i y de arcos u_j, $1 \leq i \leq n$, $1 \leq j \leq m$. La matriz de incidencia de G es $B(G) = (b_{ij})$, donde:*

$$b_{ij} = \left\{ \begin{array}{cc} 1 & \textit{si } x_i \textit{ es vértice inicial de } u_j \\ -1 & \textit{si } x_i \textit{ es vértice final de } u_j \\ 0 & \textit{caso contrario} \end{array} \right\}$$

Ejemplo 2.4.13. Sea el siguiente grafo dirigido G

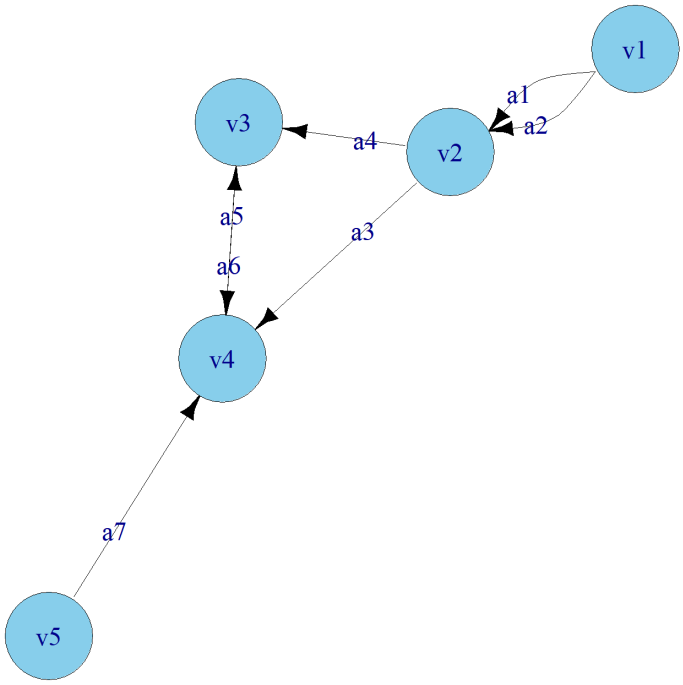

Figura 25: Grafo del ejemplo 2.4.13

Se presenta a continuación su matriz de incidencia, la misma es de orden 5×7, ya que tiene 5 vértices, representado cada uno por una fila y 7 arcos, cada uno representado por una columna:

$$B(G) = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

2.4.3 Matriz laplaciana

Existe otra matriz, estrechamente relacionada con la matriz de adyacencia pero que difiere en algunos aspectos importantes, que también puede decirnos mucho sobre la estructura de la red. Esta es la matriz laplaciana del grafo. Para introducirla, conviene comenzar con el concepto de difusión.

La difusión es, entre otras cosas, el proceso por el cual un gas se mueve desde regiones de alta densidad hacia regiones de baja densidad, impulsado por la presión relativa (o presión parcial) de las diferentes regiones (Newman, 2010).

También se pueden considerar procesos de difusión en redes, y dichos procesos se utilizan a veces como un modelo simple de propagación en una red, como la propagación de una idea o la propagación de una enfermedad.

Supongamos que tenemos alguna mercancía o sustancia de algún tipo en los vértices de una red y que hay una cantidad ψ_i de esta en el vértice i. Y supongamos que la mercancía se mueve a lo largo de las aristas, fluyendo de un vértice j a uno adyacente i a una tasa $C(\psi_j - \psi_i)$ donde C es una constante

llamada constante de difusión. Es decir, en un pequeño intervalo de tiempo, la cantidad de fluido que fluye de j a i es $C(\psi_j - \psi_i)dt$. Entonces, la tasa a la que ψ_i está cambiando está dada por:

$$\frac{d\psi_i}{dt} = C\sum_j A_{ij}(\psi_j - \psi_i)$$

(1)

La matriz de adyacencia en esta expresión asegura que los únicos términos que aparecen en la suma son aquellos que corresponden a pares de vértices que están efectivamente conectados por una arista.

La ecuación (1) funciona igualmente bien tanto para redes dirigidas como no dirigidas, pero enfoquémonos aquí en las redes no dirigidas. También consideraremos que nuestras redes son simples (es decir, que tienen como máximo una sola arista entre cada par de vértices y no tienen auto-aristas).

Separando los dos términos en la ecuación (1), podemos escribir:

$$\begin{aligned}\frac{d\psi_i}{dt} &= C\sum_j A_{ij}\psi_j - C\psi_i\sum_j A_{ij} = C\sum_j A_{ij}\psi_j - C\psi_i k_i \\ &= C\sum_j (A_{ij} - \delta_{ij}k_i)\psi_j\end{aligned}$$

(2)

donde k_i es el grado del vértice i, como es habitual, y hemos hecho uso del resultado $k_i = \sum_j A_{ij}$. (Y δ_{ij} es el delta de Kronecker, que vale 1 si $i = j$ y 0 en caso contrario.)

La ecuación (2) puede escribirse en forma matricial como:

$$\frac{d\psi}{dt} = C(\mathbf{A} - \mathbf{D})\psi$$

(3)

donde ψ es el vector cuyas componentes son los valores ψ_i, $\mathbf{A}$ es la matriz de adyacencia, y $\mathbf{D}$ es la matriz diagonal con los grados de los vértices en su diagonal:

$$\mathbf{D} = \begin{pmatrix} k_1 & 0 & 0 & ... \\ 0 & k_2 & 0 & ... \\ 0 & 0 & k_3 & ... \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \end{pmatrix}$$

(4)

Es común definir a la nueva matriz como

$$\mathbf{L} = \mathbf{D} - \mathbf{A}$$

(5)

de modo que la ecuación (2) toma la forma

$$\frac{d\psi}{dt} + C\mathbf{L}\psi = 0$$

(6)

lo cual tiene la misma forma que la ecuación de difusión ordinaria para un gas, excepto que el operador laplaciano ∇^2 que aparece en dicha ecuación ha sido reemplazado por la matriz $\mathbf{L}$.

Por esta razón, la matriz $\mathbf{L}$ se llama el laplaciano del grafo, aunque su importancia va mucho más allá de los procesos de difusión. Como veremos, el laplaciano del grafo aparece en una variedad de contextos diferentes, incluyendo caminos aleatorios en redes, redes de resistencias, partición de grafos y conectividad de redes.

Escrito en su forma completa, los elementos de la matriz laplaciana son:

$$L_{ij} = \left\{ \begin{array}{ccc} k_i & si & i = j \\ -1 & si & i \neq j \text{ y hay un vértice } (i,j) \\ 0 & & \text{de lo contrario} \end{array} \right\}$$

(7)

por lo tanto, tiene los grados de los vértices en su diagonal y un elemento -1 por cada arista.

Alternativamente, podemos escribir:

$$L_{ij} = \delta_{ij} k_i - A_{ij}$$

(8)

Podemos resolver la ecuación de difusión (6) escribiendo el vector ψ como una combinación lineal de los autovectores $\mathbf{v}_i$ del laplaciano, de la siguiente manera:

$$\psi(t) = \sum_i a_i(t) \mathbf{v}_i$$

(9)

con los coeficientes $a_i(t)$ que varían con el tiempo. Esto tiene sentido porque el laplaciano $\mathbf{L}$ del grafo es una matriz simétrica y real, por lo tanto tiene un conjunto completo de autovectores ortogonales que forman una base del espacio $\mathbb{R}^n$. Así, cualquier vector, incluido $\psi(t)$, puede descomponerse como combinación lineal de esos autovectores. Sustituyendo esta forma en la ecuación (6) y utilizando que $\mathbf{L}\mathbf{v}_i = \lambda_i \mathbf{v}_i$, donde λ_i es el autovalor correspondiente al autovector $\mathbf{v}_i$, obtenemos:

$$\begin{aligned} \frac{d\psi}{dt} + C\mathbf{L}\psi &= 0 \\ \frac{d}{dt}\sum_i a_i \mathbf{v}_i + C\mathbf{L}\sum_i a_i \mathbf{v}_i &= 0 \\ \sum_i \frac{d}{dt} a_i \mathbf{v}_i + \sum_i C a_i \mathbf{L}\mathbf{v}_i &= 0 \\ \sum_i (\frac{d}{dt} a_i \mathbf{v}_i + C\lambda_i \mathbf{v}_i a_i) &= 0 \\ \sum_i (\frac{da_i}{dt} + C\lambda_i a_i)\mathbf{v}_i &= 0 \end{aligned}$$

(10)

Pero los autovectores de una matriz simétrica como el laplaciano son ortogonales, y por lo tanto, al tomar el producto punto de esta ecuación con cualquier autovector $\mathbf{v}_j$, se eliminan todos los términos de la suma excepto aquel con $i = j$. Luego obtenemos:

$$\begin{aligned} (\frac{da_i}{dt} + C\lambda_i a_i) \left\|\mathbf{v}_i\right\|^2 &= 0 \\ \frac{da_i}{dt} + C\lambda_i a_i &= 0 \end{aligned}$$

(11)

para todo i. Nótese que $\|v_i\| = 1$, dado que por ser $\mathbf{L}$ una matriz simétrica real, sus autovectores forman una base ortonormal, así que sus vectores están normalizados. Esto último vale por el teorema espectral para matrices simétricas reales, el cual garantiza que toda matriz simétrica real es diagonalizable mediante una base ortonormal de autovectores (para una exposición más detallada, véase, por ejemplo, Strang (2009) o Poole (2010)). Por último, la ecuación diferencial tiene la siguiente solución.

$$a_i(t) = a_i(0)e^{-C\lambda_i t}$$

(12)

Dada una condición inicial para el sistema, especificada por las cantidades $a_i(0)$, podemos entonces resolver para el estado en cualquier instante posterior, siempre que conozcamos los autovalores y autovectores del Laplaciano del grafo.

Observación 2: Al conjunto de autovalores de una matriz en ocasiones se llama espectro de la matriz. Esto es debido a que espectro es una palabra latina que significa "imagen". Cuando los átomos vibran, emiten luz. Y cuando la luz pasa a través de un prisma, se dispersa en un espectro: una banda de colores del arco iris. Las frecuencias de vibración corresponden a los autovalores de cierto operador y son visibles como líneas brillantes en el espectro de luz que se emite desde un prisma. Por ende, literalmente pueden verse los autovalores del átomo en su espectro y, por esta razón, es adecuado que se aplique la palabra espectro al conjunto de todos los autovalores de una matriz (u operador en general).

Observación 3: Así como $L = D - A$, también se puede definir la matriz Q, llamada matriz laplaciana sin signo, que es igual a la suma de las matrices D y A, es decir $Q = D + A$. Una importante propiedad de las matrices L y Q es que son semidefinidas positivas (sus autovalores son no negativos). Además $Q = BB^T$ y $L = NN^T$, siendo B la matriz de incidencia de G y N la matriz de incidencia de un grafo dirigido obtenido a partir de G con cualquier orientación de sus aristas.

Ejemplo 2.4.14. Hallar la matriz laplaciana del siguiente grafo G

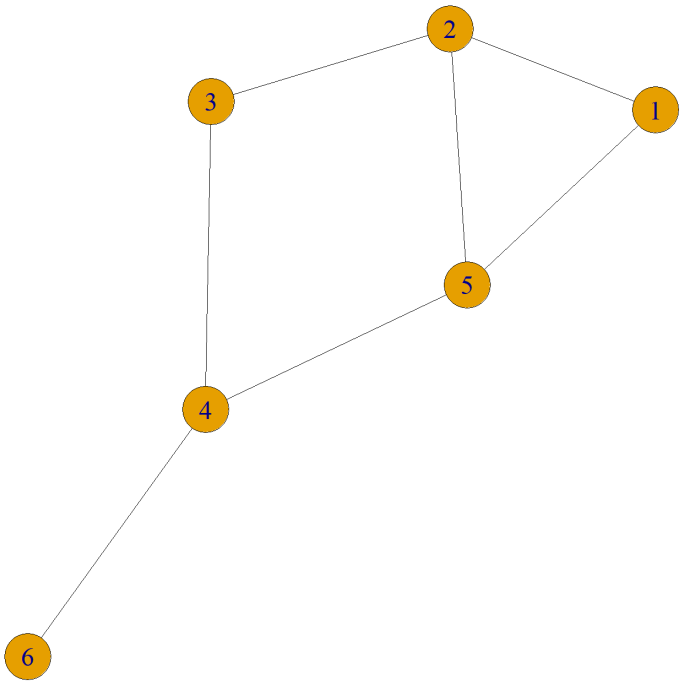

Figura 26: Grafo G del ejemplo 2.4.14

Primero calculamos la matriz de adyacencia y de grados del grafo G.
Matriz de adyacencia de G

$$A(G) = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Matriz de grados de G

$$D(G) = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Ahora como $L(G) = D(G) - A(G)$ nos queda que la matriz laplaciana de G es

$$L(G) = \begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & -1 \end{pmatrix}$$

Veamos ahora algunas propiedades de la matriz laplaciana.

La siguiente proposición resume algunas de las propiedades más importantes de L (von Luxburg, 2007).

Proposición 2.4.15 (Propiedades de L)**.** *La matriz L satisface las siguientes propiedades:*

1. Para cada vector $f \in \mathbb{R}^n$ tenemos

$$f'Lf = \frac{1}{2}\sum_{i,j=1}^{n} a_{ij}(f_i - f_j)^2$$

siendo f' el vector transpuesto de f, es decir es f^T.

2. L es simétrica y positiva semidefinida.

3. El autovalor más pequeño de L es 0 y su correspondiente autovector es el vector constante $\mathbf{1}$.

4. L tiene n autovalores reales no negativos $0 = \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$.

Dem:

1) Por la definición de $D = diag(d_1, d_2, ..., d_n)$ sabemos que $d_i = d_j$ cuando $i = j$ y que $d_i = 0$ cuando $i \neq j$, además de que $\sum_{i=1}^{n} a_{ij} = d_j$, ya que por la definición de matriz de adyacencia sabemos que la suma de los elementos de una fila (o columna) será igual al grado del nodo, es decir d_i, con esas cosas en mente se plantea:

$$\begin{aligned}
f'Lf &= f'(D-A)f \\
f'Lf &= f'Df - f'Af \\
&= \sum_{i=1}^{n} d_i f_i^2 - \sum_{i,j=1}^{n} f_i f_j a_{ij} \\
&= \frac{1}{2}(\sum_{i=1}^{n} d_i f_i^2 - 2\sum_{i,j=1}^{n} f_i f_j a_{ij} + \sum_{j=1}^{n} d_j f_j^2) \\
&= \frac{1}{2}(\sum_{i=1}^{n} f_i^2 \sum_{j=1}^{n} a_{ij} - 2\sum_{i,j=1}^{n} f_i f_j a_{ij} + \sum_{j=1}^{n} f_j^2 \sum_{i=1}^{n} a_{ij}) \\
&= \frac{1}{2}(\sum_{i,j=1}^{n} a_{ij} f_i^2 - 2\sum_{i,j=1}^{n} f_i f_j a_{ij} + \sum_{i,j=1}^{n} a_{ij} f_j^2) \\
&= \frac{1}{2}\sum_{i,j=1}^{n} a_{ij}(f_i^2 - 2f_i f_j + f_j^2) \\
&= \frac{1}{2}\sum_{i,j=1}^{n} a_{ij}(f_i - f_j)^2
\end{aligned}$$

2) L es simétrica porque D y A lo son, mientras que por 1) $f'Lf \geq 0$ para todo $f \in \mathbb{R}^n$, por lo que L es semidefinida positiva.

3) Debemos probar que $\mathbf{1}$ es un autovector de L con autovalor 0, es decir $L\mathbf{1} = 0$, pero como

$\sum_{i=1}^{n} a_{ij} = d_j$ tenemos que

$$\begin{aligned} L\mathbf{1} &= (D-A)\mathbf{1} \\ &= D\mathbf{1} - A\mathbf{1} \\ &= \begin{pmatrix} d_1 \\ d_2 \\ . \\ . \\ . \\ d_n \end{pmatrix} - \begin{pmatrix} d_1 \\ d_2 \\ . \\ . \\ . \\ d_n \end{pmatrix} \\ &= 0 \end{aligned}$$

Luego $\mathbf{1}$ es un autovector de L con autovalor 0, y como por 2) L es simétrica y semidefinida positiva sabemos que sus autovalores son reales y no negativos, así que 0 es el más pequeño de ellos.

4) Es consecuencia directa de los incisos anteriores, como L es una matriz de $n \times n$ y ya vimos que sus autovalores son reales y no negativos, y el más pequeño es 0, nos queda que $0 = \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$.

Ejemplo 2.4.16. Sea el grafo H

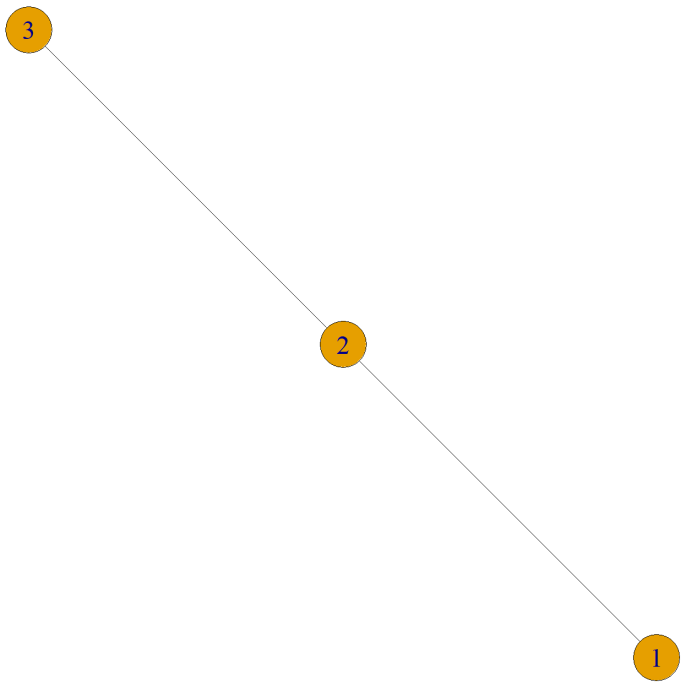

Figura 27: Grafo H

Sus matrices de adyacencia, diagonal y laplaciana respectivamente son

$$\begin{aligned} A(H) &= \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \\ D(H) &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ L(H) &= \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} \end{aligned}$$

Sea $f \in \mathbb{R}^3$, $f = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ y $f' = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$, verifiquemos que se cumple el inciso 1) de la proposición

$$\begin{aligned} f'Lf &= \begin{pmatrix} 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \\ f'Lf &= \begin{pmatrix} -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \\ f'Lf &= (-1+3) \\ f'Lf &= 2 \end{aligned}$$

Ahora veamos el otro lado de la igualdad

$$\begin{aligned} \frac{1}{2}\sum_{i,j=1}^{3} a_{ij}(f_i - f_j)^2 &= \frac{1}{2}\left[a_{12}(f_1-f_2)^2 + a_{21}(f_2-f_1)^2 + a_{23}(f_2-f_3)^2 + a_{32}(f_3-f_2)^2\right] \\ \frac{1}{2}\sum_{i,j=1}^{3} a_{ij}(f_i - f_j)^2 &= \frac{1}{2}\left[(1-2)^2 + (2-1)^2 + (2-3)^2 + (3-2)^2\right] \\ \frac{1}{2}\sum_{i,j=1}^{3} a_{ij}(f_i - f_j)^2 &= \frac{1}{2}\left[1+1+1+1\right] \\ \frac{1}{2}\sum_{i,j=1}^{3} a_{ij}(f_i - f_j)^2 &= \frac{1}{2}4 \\ \frac{1}{2}\sum_{i,j=1}^{3} a_{ij}(f_i - f_j)^2 &= 2 \end{aligned}$$

Por lo tanto se cumple que $f'Lf = \frac{1}{2}\sum_{i,j=1}^{n} a_{ij}(f_i - f_j)^2$ así que se verifica el inciso 1) de la proposición.
Ahora calculemos los autovectores de la matriz laplaciana

$$\begin{aligned} \det(L-\lambda I) &= \begin{pmatrix} 1-\lambda & -1 & 0 \\ -1 & 2-\lambda & -1 \\ 0 & -1 & 1-\lambda \end{pmatrix} \\ \det(L-\lambda I) &= -\lambda(\lambda-1)(\lambda-3) = 0 \end{aligned}$$

Tenemos que los autovalores son $\lambda_1 = 0, \lambda_2 = 1$ y $\lambda_3 = 3$. Como son todos reales no negativos se verifica 4). Esto significa que L es positiva y semi-definida, además al ser L claramente simétrica también se verifica 2). Por último, veamos que el autovector asociado al autovalor 0 (que es el más pequeño de L) es el autovector constante $\mathbf{1}$.

$$\begin{aligned} L\mathbf{1} &= \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \\ L\mathbf{1} &= \begin{pmatrix} 1-1+0 \\ -1+2-1 \\ 0-1+1 \end{pmatrix} \\ L\mathbf{1} &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \end{aligned}$$

De esta forma queda comprobado 3) y así el ejemplo verifica la proposición completa.

El laplaciano de grafos y sus autovalores y autovectores pueden utilizarse para describir numerosas propiedades de grafos. Un ejemplo que será importante para la agrupación espectral es el siguiente:

Proposición 2.4.17 (Número de componentes conexas y el espectro de L)**.** *Sea G un grafo no dirigido con pesos no negativos. Entonces, la multiplicidad k del autovalor 0 de la matriz laplaciana L es igual al número de componentes conexas $A_1, ..., A_k$ del grafo. El espacio propio asociado al autovalor 0 está generado por los vectores indicadores $\mathbf{1}_{A_1}, ..., \mathbf{1}_{A_k}$ de esas componentes.*

Dem: Comenzamos con el caso $k = 1$, es decir, el grafo es conexo. Supongamos que f es un autovector con su autovalor asociado 0. Entonces, sabemos que

$$0 = f'Lf = \sum_{i,j=1}^{n} a_{ij}(f_i - f_j)^2$$

Dado que los pesos a_{ij} son no negativos, esta suma solo puede anularse si todos los términos $a_{ij}(f_i - f_j)^2$ se anulan. Por lo tanto, si dos vértices v_i y v_j están conectados (es decir, $a_{ij} > 0$), entonces f_i debe ser igual a f_j. Con este argumento, podemos ver que f debe ser constante para todos los vértices que puedan conectarse mediante un camino en el grafo. Además, como todos los vértices de una componente conexa en un grafo no dirigido pueden conectarse mediante un camino, f debe ser constante en toda la componente conexa. En un grafo que consiste en una única componente conexa, entonces, solo tenemos el vector constante $\mathbf{1}$ como vector propio con valor propio 0, que obviamente es el vector indicador de la componente conexa.

Ahora consideremos el caso de k componentes conexas. Sin pérdida de generalidad, asumimos que los vértices están ordenados de acuerdo con las componentes conexas a las que pertenecen. En este caso, la matriz de adyacencia A tiene una forma de bloque diagonal, y lo mismo ocurre con la matriz L:

$$L = \begin{pmatrix} L_1 & & & \\ & L_2 & & \\ & & ... & \\ & & & L_k \end{pmatrix}$$

Nótese que cada uno de los bloques L_i es en sí mismo una matriz laplaciana propiamente dicha, es decir, el laplaciano correspondiente al subgrafo de la i-ésima componente conexa. Como ocurre con todas las matrices en forma de bloque diagonal, sabemos que el espectro de L está dado por la

unión de los espectros de los L_i, y que los autovectores correspondientes de L son los autovectores de los L_i, rellenados con 0 en las posiciones correspondientes a los otros bloques. Como cada L_i es la matriz laplaciana de un grafo conexo, sabemos que cada L_i tiene el autovalor 0 con multiplicidad 1, y el autovector correspondiente es el vector constante $\mathbf{1}$ uno sobre la i-ésima componente conexa. Por lo tanto, la matriz L tiene tantos autovalores 0 como componentes conexas hay, y los autovectores correspondientes son los vectores indicadores de las componentes conexas.

Ejemplo 2.4.18. Consideremos el siguiente grafo G

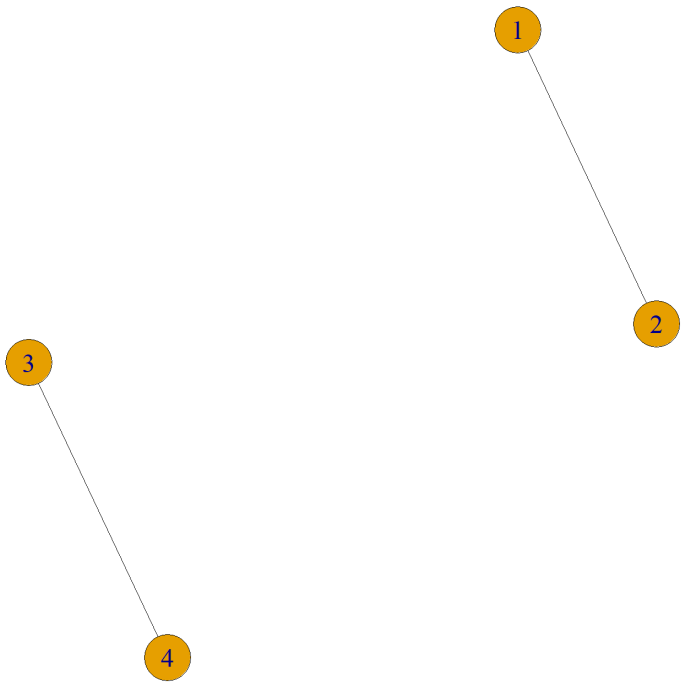

Figura 28: Grafo G con dos componentes conexas

Este grafo tiene 4 vértice y dos componentes conexas: componente A_1 con los vértice 1 y 2 conec-

tados y componente A_2 con los vértices 3 y 4 conectados. Veamos como son sus diferentes matrices

$$
\begin{aligned}
A(G) &= \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \\
D(G) &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \\
L(G) &= \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}
\end{aligned}
$$

Si calculamos los autovalores de L obtendremos que son $\lambda = \{0, 0, 2, 2\}$. Esto significa que el autovalor 0 tiene multiplicidad 2, lo que coincide con la cantidad de componentes conexas del grafo, cumpliendose así la primera parte de la proposición, ya que este grafo tiene dos componentes conexas y entonces el autovalor 0 de la matriz laplaciana tiene multiplicidad 2. Luego, los autovectores asociados a 0 son $v_1 = (1, 1, 0, 0)^T$ y $v_2 = (0, 0, 1, 1)^T$, los cuales podemos notar que son los vectores indicadores de las componentes A_1 y A_2 respectivamente, en otras palabras, los vectores propios correspondientes a 0 son los que identifican a cada componente, además de que el espacio propio de $\lambda = 0$ está generado por los vectores $\mathbf{1}_{A_1}, \mathbf{1}_{A_2}$. De esta forma queda verificada la proposición en este ejemplo.

2.5 Detección de comunidades

En esta sección examinaremos el problema de la detección de comunidades, que consiste en buscar los grupos que se forman de manera natural en una red, sin importar su número o tamaño. Esta técnica se utiliza principalmente como una herramienta para descubrir y comprender la estructura a gran escala de las redes (Newman, 2010).

El objetivo básico de la detección de comunidades es el siguiente: queremos separar la red en grupos de vértices que tengan pocas conexiones entre ellos, con el número y el tamaño de los grupos sin estar fijados. Comencemos centrándonos en un ejemplo muy simple de un problema de detección de comunidades, probablemente el más simple. Consideraremos el problema de dividir una red en solo dos grupos o comunidades que no se superpongan, es decir no solapados, sin ninguna restricción sobre los tamaños de los grupos, salvo que la suma de los tamaños debe ser igual al tamaño total n de la red. Así, en esta versión simple del problema, el número de grupos está determinado, pero sus tamaños no, y deseamos encontrar la división "natural" de la red en dos grupos: la línea de falla (si existe) a lo largo de la cual la red se divide de forma inherente, aunque aún no hemos definido con precisión qué queremos decir con eso, por lo que la pregunta que estamos planteando todavía no está bien definida.

Nuestra primera suposición sobre cómo abordar este problema podría ser simplemente encontrar la división con el tamaño mínimo de corte, pero sin ninguna restricción sobre los tamaños de los grupos. Sin embargo, basta una breve reflexión para ver que esto no funcionará. Si dividimos una red en dos grupos permitiendo cualquier cantidad de vértices en ellos, entonces la división óptima simplemente consiste en poner todos los vértices en uno de los grupos y ninguno en el otro. Esta división trivial asegura que el tamaño del corte entre los dos grupos será cero—no habrá aristas entre grupos porque uno de ellos no contiene vértices. Como respuesta a nuestro problema de detección de comunidades, sin embargo, esta solución claramente no es útil.

Una forma de mejorar esto sería imponer restricciones flexibles sobre los tamaños de los grupos. Es decir, podríamos permitir que los tamaños de los grupos varíen, pero no demasiado. Un ejemplo

de este tipo de enfoque es la partición por corte de razón (ratio cut), en la cual, en lugar de minimizar el corte estándar R, se minimiza el cociente $R/(n_1 n_2)$, donde n_1 y n_2 son los tamaños de los dos grupos. El denominador $n_1 n_2$ alcanza su valor más grande, y por lo tanto reduce el cociente en la mayor cantidad posible, cuando n_1 y n_2 son iguales, o sea $n_1 = n_2 = \frac{1}{2}n$. Para tamaños desiguales, el denominador disminuye cuanto mayor es la desigualdad, y tiende a cero cuando alguno de los grupos tiene tamaño cero. Esto elimina efectivamente las soluciones en las que todos los vértices se colocan en el mismo grupo, ya que tales soluciones nunca dan el valor mínimo del cociente, y favorece divisiones en las que los grupos tienen tamaños aproximadamente iguales.

Sin embargo, como herramienta para descubrir las divisiones naturales en una red, el ratio cut no es ideal. En particular, aunque permite que los tamaños de los grupos varíen, sigue estando sesgado hacia una elección específica, la de grupos de igual tamaño. Más importante aún, no hay un fundamento teórico detrás de su definición. Funciona razonablemente bien en algunas circunstancias, pero no hay ninguna razón fundamental para creer que dará respuestas sensatas o que algún otro enfoque no dará mejores resultados.

Una estrategia alternativa consiste en centrarse en una medida diferente de la calidad de una división, más allá del simple tamaño del corte o sus variantes. Se ha argumentado que el tamaño del corte no es en sí una buena medida, porque una buena división de una red en comunidades no es solamente aquella en la que hay pocas aristas entre comunidades. Por el contrario, se argumenta que una buena división es aquella en la que hay menos aristas de las esperadas entre los grupos. Si encontramos una división de una red que tiene pocas aristas entre sus grupos, pero aun así el número de esas aristas es aproximadamente el que esperaríamos si las aristas se colocaran aleatoriamente en la red, entonces la mayoría de las personas diría que no hemos encontrado nada significativo. No es el tamaño total del corte lo que importa, sino cómo se compara ese tamaño con lo que esperamos observar.

De hecho, en el desarrollo convencional de esta idea no se considera el número de aristas entre grupos, sino el número de aristas dentro de los grupos. Sin embargo, ambos enfoques son equivalentes, ya que toda arista que está dentro de un grupo no necesariamente está entre grupos, así que uno puede calcular un valor a partir del otro, dado el número total de aristas en la red en su conjunto. Nosotros seguiremos la convención aquí y basaremos nuestros cálculos en el número de aristas dentro de los grupos.

Nuestro objetivo, por tanto, será encontrar una medida que cuantifique cuántas aristas hay dentro de los grupos en nuestra red, en relación con la cantidad de aristas que esperaríamos encontrar allí por azar. Sin embargo, esta es una idea que ya ha sido considerada en otros contextos. Por ejemplo, en el estudio del fenómeno de mezcla asortativa en redes, donde los vértices con características similares tienden a conectarse entre sí, se ha introducido una medida conocida como modularidad, que toma un valor alto cuando hay muchas más aristas entre vértices del mismo tipo de lo que se esperaría por azar. Este es precisamente el tipo de medida que necesitamos para resolver nuestro problema actual de detección de comunidades. Si consideramos que los vértices de nuestros dos grupos representan vértices de dos tipos distintos, entonces las buenas divisiones de la red en comunidades serán precisamente aquellas que tengan altos valores de modularidad correspondiente.

Así, una forma de detectar comunidades en redes es buscar las divisiones que tengan los mayores valores de modularidad, y de hecho, este es el método más comúnmente utilizado para la detección de comunidades. La maximización de la modularidad es un problema difícil. Se cree que, los únicos algoritmos capaces de encontrar siempre la división con máxima modularidad tardan un tiempo exponencial en ejecutarse, y por tanto son inútiles para redes que no sean muy pequeñas. Por ello, se recurren a algoritmos heurísticos, es decir, algoritmos que intentan maximizar la modularidad de forma inteligente, obteniendo resultados razonablemente buenos en la mayoría de los casos. No obstante, no existe una definición universalmente aceptada de lo que constituye una buena división de una red en comunidades. Si bien algunos algoritmos se basan en una formulación específica mediante la función de modularidad, existen diversas definiciones alternativas que han dado lugar a métodos distintos. En los siguientes capítulos se presentarán algunos de estos enfoques, que no recurren directamente a la optimización de la modularidad.

2.5.1 Intermediación de aristas

Uno de los posibles enfoques para la detección de comunidades es aquel centrado en las aristas. En lugar de intentar construir una medida que nos diga qué aristas son más centrales dentro de las comunidades, nos enfocamos en aquellas aristas que son menos centrales, es decir, las aristas que están más "entre" comunidades. En lugar de construir comunidades agregando las aristas más fuertes a un conjunto de vértices inicialmente vacío, las construimos eliminando progresivamente aristas del grafo original.

La intermediación de vértices ha sido estudiada en el pasado como una medida de la centralidad e influencia de los nodos en las redes. Propuesta por primera vez por Freeman, la centralidad de intermediación de un vértice i se define como el número de caminos más cortos entre pares de otros vértices que pasan por i. Es una medida de la influencia de un nodo sobre el flujo de información entre otros nodos, especialmente en casos donde el flujo de información en una red sigue principalmente los caminos más cortos disponibles.

Para encontrar qué aristas en una red están más "entre" otros pares de vértices, generalizamos la centralidad de intermediación de Freeman a las aristas y definimos la **intermediación de una arista** como el número de caminos más cortos entre pares de vértices que pasan por ella (Girvan & Newman, 2002). Si hay más de un camino más corto entre un par de vértices, cada camino recibe el mismo peso de modo que el peso total de todos los caminos es uno.

Si una red contiene comunidades o grupos que están conectados débilmente entre sí por unas pocas aristas intergrupales, entonces todos los caminos más cortos entre diferentes comunidades deben pasar por una de estas pocas aristas. Por lo tanto, las aristas que conectan comunidades tendrán una alta intermediación. Al eliminar estas aristas, separamos los grupos entre sí y revelamos así la estructura de comunidades subyacente del grafo.

El algoritmo que proponemos para identificar comunidades se enuncia de la siguiente manera:

1. Calcular la intermediación de todas las aristas en la red.
2. Eliminar la arista con mayor intermediación.
3. Recalcular las intermediaciones de todas las aristas afectadas por la eliminación.
4. Repetir desde el paso 2 hasta que no queden aristas.

Desde el punto de vista práctico, calculamos las intermediaciones utilizando el algoritmo rápido de Newman, que calcula la intermediación para las m aristas de un grafo con n vértices en un tiempo $O(mn)$ (es decir que el tiempo de ejecución crece proporcionalmente al producto del número de aristas por el número de vértices). Dado que este cálculo debe repetirse tras la eliminación de cada arista, el algoritmo completo se ejecuta, en el peor de los casos, en tiempo $O(m^2n)$.

Sin embargo, después de eliminar cada arista, solo es necesario recalcular las intermediaciones de aquellas aristas que se vieron afectadas por la eliminación, lo cual, como mucho, incluye solo las que están en el mismo componente que la arista eliminada. Esto significa que el tiempo de ejecución puede ser mejor que el peor caso en redes con una estructura de comunidades fuerte (es decir, aquellas que rápidamente se fragmentan en componentes separados tras las primeras iteraciones del algoritmo).

Para tratar de reducir aún más el tiempo de ejecución del algoritmo, uno podría verse tentado a calcular las intermediaciones de todas las aristas una sola vez y luego eliminarlas en orden decreciente según su intermediación. Sin embargo, descubrimos que esta estrategia no funciona bien, ya que si dos comunidades están conectadas por más de una arista, no hay garantía de que todas esas aristas tengan una alta intermediación: solo sabemos que al menos una la tendrá. Al recalcular las intermediaciones tras la eliminación de cada arista, aseguramos que al menos una de las aristas restantes entre dos comunidades siempre tendrá un valor alto.

Ejemplo 2.5.1. Identificar las comunidades del siguiente grafo G utilizando el algoritmo de Girvan-Newman.

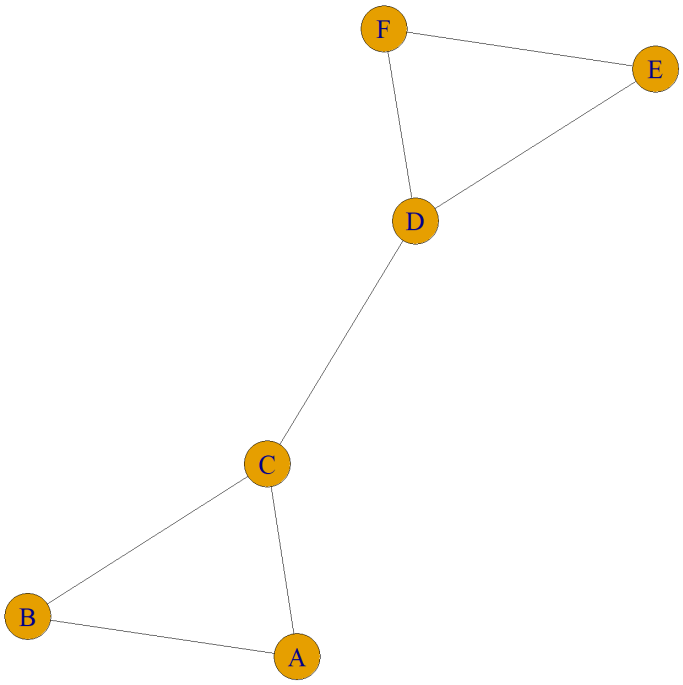

Figura 29: Grafo G del ejemplo 2.5.1

Calculamos la intermediación de todas las aristas de la red, contando primero los caminos más cortos entre pares de vértices que pasan por las aristas.

Caminos más cortos desde	Camino
A-B	$A \to B$
A-C	$A \to C$
A-D	$A \to C \to D$
A-E	$A \to C \to D \to E$
A-F	$A \to C \to D \to F$
B-C	$B \to C$
B-D	$B \to C \to D$
B-E	$B \to C \to D \to E$
B-F	$B \to C \to D \to F$
C-D	$C \to D$
C-E	$C \to D \to E$
C-F	$C \to D \to F$
D-E	$D \to E$
D-F	$D \to F$
E-F	$E \to F$

Conociendo los caminos tenemos que la intermediación de aristas es

$$\begin{pmatrix} \text{Arista} & \text{Intermediación} \\ \text{A} \to \text{B} & 1 \\ \text{A} \to \text{C} & 4 \\ \text{B} \to \text{C} & 4 \\ \text{C} \to \text{D} & 9 \\ \text{D} \to \text{E} & 4 \\ \text{D} \to \text{F} & 4 \\ \text{E} \to \text{F} & 1 \end{pmatrix}$$

Como la arista C→D es la que tiene el valor de intermediación más alto la eliminamos, luego el grafo queda.

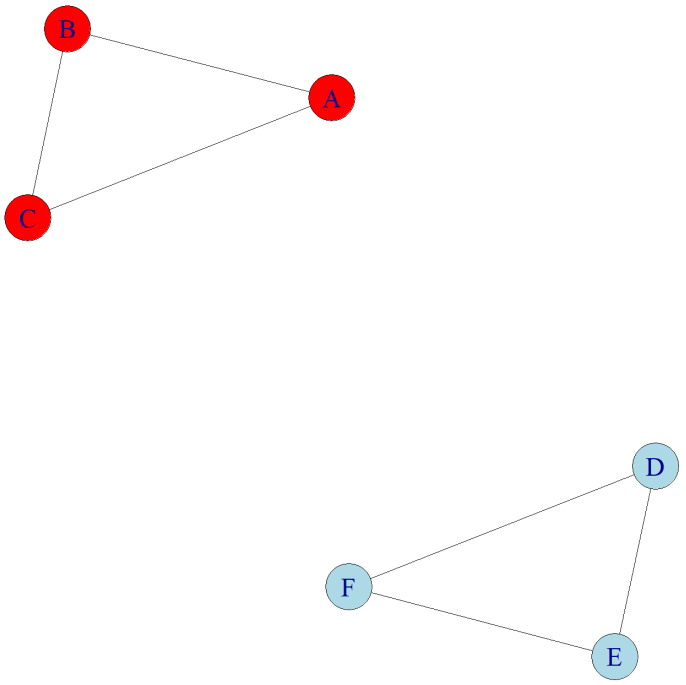

Figura 30: Grafo G con comunidades identificadas

Claramente ahora las aristas restantes tienen todas la misma intermediación, además que se pueden identificar fácilmente dos comunidades, una con los nodos A,B y C, y una segunda comunidad con los nodos D,E y F.

2.5.2 Agrupamiento particionado: algoritmo k-means

Uno de los enfoques posibles para la detección de comunidades es el agrupamiento (o clustering) particionado, que busca dividir el conjunto de nodos en un número fijo de grupos, maximizando la cohesión interna y minimizando la dispersión entre grupos. Un algoritmo representativo de esta clase

es k-means, ampliamente utilizado en distintos dominios del análisis de datos. Aunque su aplicación directa sobre grafos no es inmediata, puede utilizarse cuando los nodos se representan en un espacio vectorial-por ejemplo, como ocurre en el agrupamiento espectral, donde se proyectan sobre autovectores seleccionados de la matriz laplaciana-. En esta sección se presentará k-means como técnica general de agrupamiento, y más adelante se aplicará dentro del marco del agrupamiento espectral.

Los algoritmos de clústering particionado construyen grupos en los que cada individuo pertenece a solo uno de ellos y el número de clústeres es solo uno (Alonso, Largo & Hoyos, 2025). Cada individuo puede pertenecer a múltiples grupos, dependiendo del punto de la construcción del algoritmo; en este caso los clústeres se sobreponen.

En este capítulo nos concentraremos en desarrollar un tipo de clústering particionado, que es intuitivo y computacionalmente no muy costoso. El clústering particionado es una forma de conformar los grupos que son mutuamente excluyentes; es decir, subconjuntos de individuos que no se sobreponen, como se presenta en la Figura 27. Estudiaremos el algoritmo k-means, que es un algoritmo de aprendizaje no supervisado, es decir, un tipo de aprendizaje automático[1] en el que el modelo intenta identificar patrones o estructuras en los datos sin que se le proporcionen etiquetas o respuestas correctas de antemano (Murphy, 2012). Con seguridad, este es uno de los algoritmos más populares y, a su vez, sencillo.

Como cualquier otra técnica de clústering, la idea principal detrás del algoritmo de k-means es dividir los individuos en k grupos de tal manera que los puntos dentro de cada grupo sean lo más similares posible entre sí y lo más diferentes posible de los puntos de otros grupos. La peculiaridad que llama la atención de esta aproximación es que va construyendo iterativamente los grupos al minimizar la suma de las distancias al cuadrado de cada punto al centroide más cercano. Intuitivamente, el centroide es un punto medio de todos los individuos que pertenecen a un grupo (ver Figura 27).

[1] El aprendizaje automático (o machine learning) puede definirse como un conjunto de métodos que detectan automáticamente patrones en los datos y los utilizan para predecir información futura o tomar otro tipo de decisiones en condiciones de incertidumbre (como planificar como recopilar más datos) (Murphy, 2012).

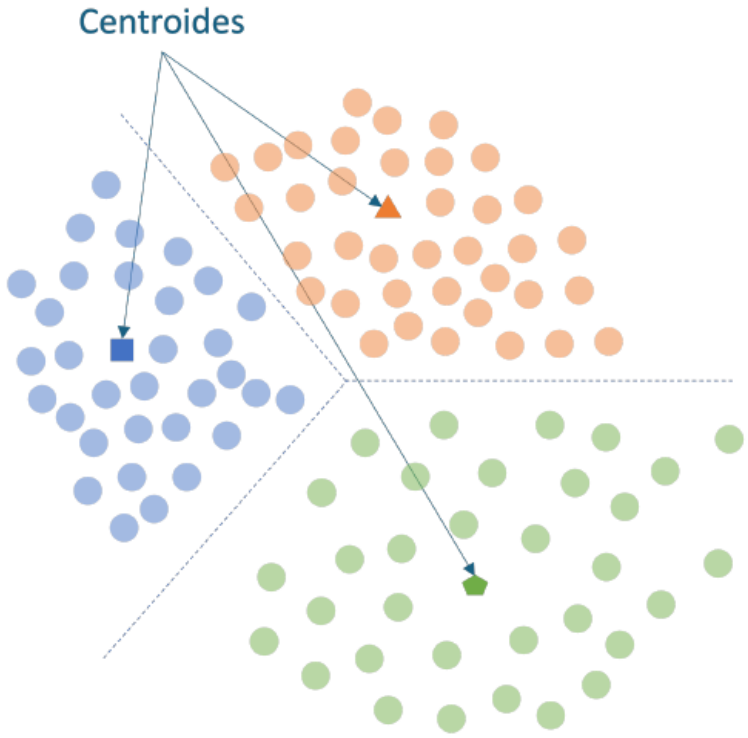

Figura 31: Representación de un algoritmo de clustering empleando k-means

Para entender la intuición de este algoritmo, veamos los pasos para realizar un clústering particionado, usando el algoritmo k-means para un determinado número de clústeres k:

1. Elegir aleatoriamente k centroides a partir de los datos disponibles.
2. Calcular las distancias de cada observación a los centroides.
3. Agrupar individuos al centroide más cercano.
4. Recalcular el centroide usando la media aritmética.
5. Agrupar de nuevo al centroide más cercano.
6. Repetir los pasos 4 y 5 hasta que no se produzca ningún cambio en la asignación de los individuos a los clústeres o se alcance un número máximo de iteraciones.

El algoritmo k-means, solo funciona para variables cuantitativas. Además, para no ver afectado el análisis de clústering por la escala de las variables, se emplean variables estandarizadas.

En la Figura 28 podemos ver el algoritmo k-means en acción. En el panel **a** tenemos k centroides al azar, en el panel **b** se calcula la distancia de cada individuo a los centroides. Luego en el panel **c** aparecen los clústeres conformados para que ya en el panel **d** se puedan ver las comunidades identificadas. Más en detalle, el algoritmo de k-means inicia seleccionando k centroides al azar, donde $k = 3$ es el número de grupos que queremos conformar. Cada color es un centroide elegido al azar.

Luego se calcula la distancia de cada individuo a los centroides y se asigna al centroide más cercano, conformando los clústeres.

A continuación, se calcula la media aritmética al interior de cada clúster, siendo esta media el nuevo centroide. Se repite el respectivo cálculo de distancia al nuevo centroide, así como la asignación de cada observación al grupo.

Esto se desarrolla de manera iterativa hasta que ya no se cambien de grupo los individuos o hasta cierto número de iteraciones que se definan arbitrariamente. Al final, tal y como se dijo antes, el objetivo es llegar a algo como lo que se presenta en el panel **d** de la Figura 28.

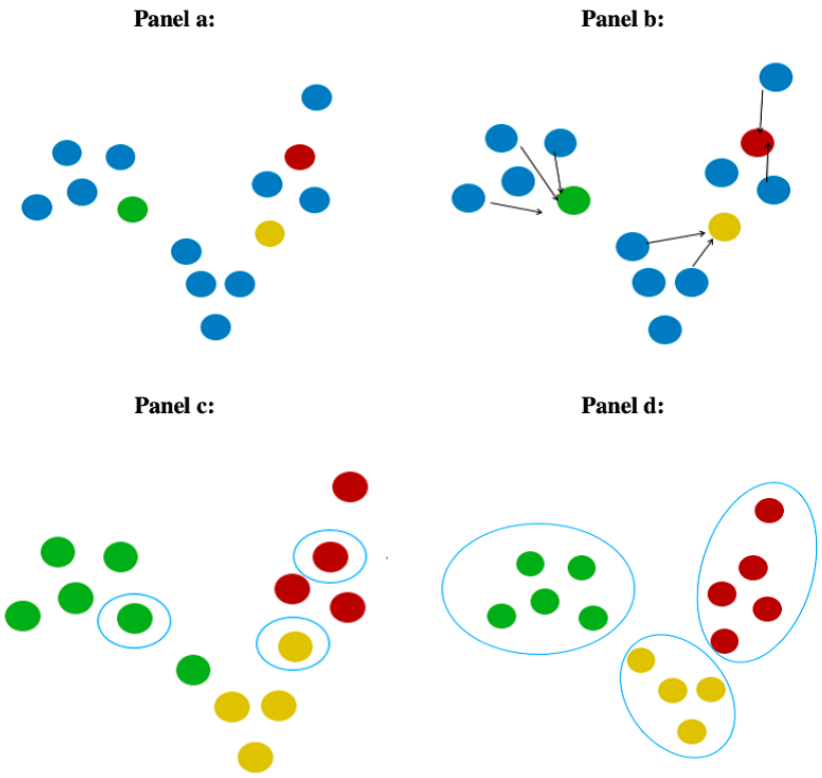

Figura 32: Algoritmo k-means. Los paneles se asocian a distintas etapas del algoritmo

Matemáticamente, para asignar cada individuo a su centroide más cercano se quiere minimizar la suma de las distancias cuadráticas a las medias. Formalmente,

$$\min_{S} \sum_{t=1}^{k} \sum_{x_j \in S_i} \|x_j - \bar{x}_i\|^2$$

(13)

donde $(x_1, x_2, x_3, ...x_n)$ son las observaciones de d dimensiones, k es el número de grupos a formar y $S = S_1, S_2, .., S_k$ es la suma de los cuadrados dentro de cada grupo. Para actualizar el centroide en la siguiente iteración $(t+1)$ se emplea la siguiente expresión:

$$\bar{x}_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j$$

(14)

El algoritmo termina cuando los individuos no cambian de clúster entre iteraciones.

2.5.3 Método del codo

Para la construcción de clústeres es necesario un criterio para determinar el número de grupos razonables, ya que seleccionar el número necesario no es una tarea trivial. Uno de los criterios más populares para determinar el número de clústeres es el método del codo, así que antes de proceder con el agrupamiento espectral, veamoslo un poco más en detalle.

Este método es tal vez el más conocido y parte de la idea fundamental de las tareas de armar agrupaciones. Recordemos que lo que se busca es encontrar clústeres, de forma que la distancia al

interior de los grupos sea la más baja posible. Es decir, que la distancia intraclúster sea la menor posible. Una medida de esa distancia intraclúster es la variación total intraclúster (WSS por su sigla en inglés de Withincluster Sum of Square) que se define como la suma de las distancias de cada una de las observaciones al respectivo centro o centroide del clúster.

La suma de cuadrados de todas las observaciones que pertenecen a un conglomerado es una medida de la variabilidad de las observaciones dentro de cada clúster. En general, un conglomerado que tiene una suma de cuadrados pequeña es más compacto que uno que tiene una suma de cuadrados grande. Así, la suma (total) de todas estas medidas de variabilidad de los clústeres (WSS) es una medida de qué tanta variabilidad presentan las observaciones para un determinado número de clústeres k.

Entonces, para encontrar el número óptimo de k, se acostumbra graficar en el eje horizontal el número de clústeres y en el eje vertical el WSS (ver Figura 29). El número óptimo de clústeres corresponderá al valor de k para el cuál se presente una caída repentina y más grande en el WSS, de tal manera que la gráfica parece tener un "codo" (de ahí el nombre del método).

En la Figura 29 se presentan diferentes representaciones de la curva WSS y la selección del número óptimo de clústeres empleando la técnica del codo. Noten que estamos buscando dónde se presente la mayor caída en el WSS de tal manera que se observa una forma de codo recogido. Por la manera como se selecciona el mejor clúster, este método de selección del número de grupos no permite comparar entre diferentes algoritmos de formación de clústeres.

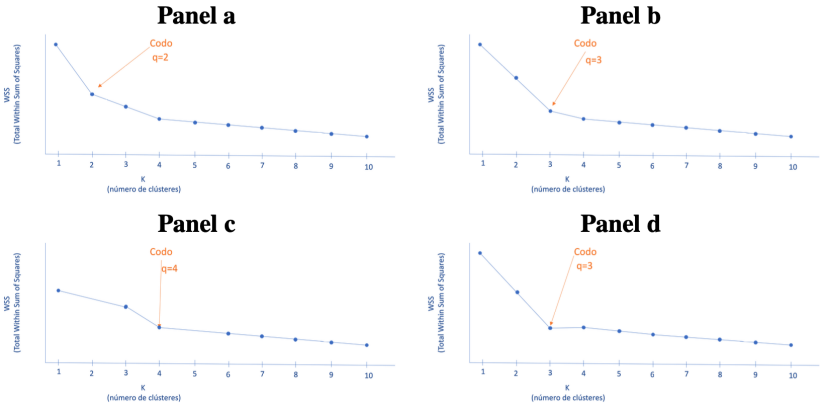

Figura 33: Representación de la selección del número de clústeres óptimos con el método del codo

2.5.4 Agrupamiento espectral

Supongamos que tenemos un conjunto de n objetos $x_1, x_2, ..., x_n$ con una función de similitud por pares sim definida entre ellos, la cual es simétrica y no negativa; es decir, $sim(x_i, x_j) = sim(x_j, x_i) \geq 0$, para todo $i, j = 1, ..., n$. A partir de esta función, es posible construir una matriz $S \in \mathbb{R}^{n \times n}$, la cual cuantifica el grado de semejanza entre cada par de elementos del conjunto. Formalmente, la entrada S_{ij} de la matriz está dada por $S_{ij} = sim(x_i, x_j)$, donde sim es la función de similitud mencionada antes que asigna un valor real no negativo a cada par de objetos. A esta matriz la llamaremos matriz de similitud S.

El agrupamiento espectral incluye todos los métodos y técnicas que particionan el conjunto en clústeres utilizando los autovectores de matrices como la matriz de similitud S u otras matrices derivadas de ella (Fortunato, 2010). En particular, los objetos podrían ser puntos en algún espacio métrico o los vértices de un grafo. El agrupamiento espectral consiste en una transformación del conjunto inicial de objetos a un conjunto de puntos en un espacio, cuyas coordenadas son elementos de autovectores: luego, este conjunto de puntos se agrupa usando técnicas estándar como el agrupamiento k-means. Uno podría preguntarse por qué es necesario agrupar los puntos obtenidos a través de los autovectores si uno podría agrupar directamente el conjunto inicial de objetos usando la matriz

de similitud. La razón es que el cambio de representación inducido por los autovectores hace que las propiedades de agrupamiento del conjunto de datos inicial sean mucho más evidentes. De este modo, el agrupamiento espectral puede separar puntos de datos que no podrían distinguirse aplicando directamente k-means, ya que este último tiende a producir conjuntos convexos de puntos.

En este capítulo seguiremos, según von Luxburg (2007), un enfoque basado en el agrupamiento espectral de grafos.

El laplaciano es, por mucho, la matriz más utilizada en agrupamiento espectral. En la sección anterior se muestra que el laplaciano no normalizado de un grafo con k componentes conexas tiene k valores propios iguales a cero. En ese caso, el laplaciano puede escribirse en forma de bloque diagonal, es decir, los vértices pueden ordenarse de tal manera que el laplaciano muestre k bloques cuadrados a lo largo de la diagonal con algunas entradas distintas de cero, mientras que todos los demás elementos son nulos. Cada bloque es el laplaciano del subgrafo correspondiente, por lo tanto, tiene como autovector trivial un vector con componentes $(1, 1, ..., 1, 1)$. Así, hay k autovectores degenerados con componentes no nulas en los vértices de un bloque, mientras que todos los demás componentes son cero. De esta manera, a partir de los componentes de los autovectores, uno puede identificar las componentes conexas del grafo. Por ejemplo, consideremos la matriz $n \times k$ cuyas columnas son estos k autovectores mencionados. La fila i-ésima de esta matriz es un vector con k componentes que representa al vértice i del grafo. Los vectores que representan a vértices en la misma componente conexa del grafo coinciden, y sus extremos se ubican sobre uno de los ejes de un sistema de coordenadas de k dimensiones (es decir, son vectores del tipo $(0, 0, ..., 1, 0, ..., 0, 0)$). Así, al graficar los vectores de los vértices, se observarán k puntos distintos, cada uno en un eje diferente, correspondientes a las componentes del grafo.

Si el grafo es conexo, pero consiste en k subgrafos débilmente conectados entre sí, el espectro del laplaciano tendrá un valor propio cero y los demás serán positivos. En este caso, el laplaciano no puede ponerse en forma de bloque diagonal: incluso si uno enumera los vértices según su pertenencia a clústeres (primero los de un clúster, luego los de otro, etc.), siempre habrá entradas no nulas fuera de los bloques. Sin embargo, los $k - 1$ valores propios no nulos más pequeños siguen estando cerca de cero, y los vectores de vértices correspondientes a los primeros k autovectores aún deberían permitir distinguir claramente los clústeres en un espacio de k dimensiones. Los vectores de vértices correspondientes al mismo clúster ya no coinciden exactamente, pero siguen estando bastante cerca entre sí. Por lo tanto, en lugar de k puntos, se observarán k grupos de puntos, con los puntos de cada grupo localizados cerca unos de otros y alejados de los demás grupos. Técnicas como k-means pueden entonces recuperar fácilmente los clústeres.

En principio, todas las matrices simétricas que pueden ponerse en forma de bloque diagonal tienen un conjunto de autovectores (tantos como bloques) cuyas entradas son distintas de cero en los vértices del bloque y cero en los demás, al igual que el laplaciano. La matriz de adyacencia también tiene esta propiedad, por ejemplo. Esta es una condición necesaria para que los autovectores se puedan usar exitosamente en el agrupamiento de grafos, pero no es suficiente. Vale la pena aclarar que si la matriz no fuera simétrica, todo lo anterior no valdría, los autovalores podrían ser complejos, o los autovectores no ser ortogonales por ejemplo. Para el caso del agrupamiento espectral, esto representa un problema porque la técnica depende de que los autovectores "reflejen" la estructura de clústeres del grafo, algo que requiere de la simetría para que funcione, por lo tanto, si la matriz no es simétrica, no se garantiza que se pueda usar para el agrupamiento espectral de forma confiable.

En el caso del laplaciano, sabemos que los autovectores correspondientes a los k valores propios más pequeños provienen cada uno de una de las componentes. En cambio, en la matriz de adyacencia $\mathbf{A}$, puede suceder que los valores propios grandes correspondan a la misma componente. Por lo tanto, si uno toma los autovectores correspondientes a los k valores propios más grandes (que son la contraparte de los valores propios bajos del laplaciano, ya que $\mathbf{L} = \mathbf{D} - \mathbf{A}$), algunos componentes estarán sobrerrepresentadas mientras que otras no aparecerán. Por eso, usar los autovectores de $\mathbf{A}$ en agrupamiento espectral no es, en general, confiable.

Además, los elementos de los autovectores correspondientes a las componentes deben estar suficientemente alejados de cero. Para entender esto, supongamos que tomamos una matriz (simétrica y en bloque diagonal), y que uno o más elementos de un autovector correspondiente a una componente

conexa están muy cerca de cero. Si se perturba el grafo, o en otras palabras, se modifica ligeramente su estructura (ocasionando algún cambio a la matriz asociada del grafo), agregando aristas entre diferentes componentes, todas las entradas de los autovectores perturbados se vuelven no nulas, y algunas pueden tener valores comparables a los elementos más pequeños de los autovectores en los bloques. Por lo tanto, distinguir los vértices de diferentes componentes puede volverse problemático, incluso si la perturbación es bastante pequeña, y las malas clasificaciones son probables. Ahora que hemos explicado por qué la matriz laplaciana es particularmente adecuada para el agrupamiento espectral, procedemos a describir el método.

El agrupamiento (clustering) espectral utiliza la matriz laplaciana **L**. Las entradas son la matriz de adyacencia **A** y el número k de clústeres que se desean recuperar. El primer paso consiste en calcular los vectores propios correspondientes a los k valores propios más pequeños de **L**. Luego, se construye la matriz **U** de tamaño $n \times k$, cuyas columnas son esos k vectores propios. Las n filas de **U** se utilizan para representar los vértices del grafo en un espacio euclidiano de k dimensiones, mediante un sistema de coordenadas cartesianas. Finalmente, los puntos son agrupados en k clústeres utilizando k-means.

En conclusión, así quedan resumidos los pasos del algoritmo de agrupamiento espectral.

Suponemos que nuestros datos consisten en n "puntos" $x_1, ..., x_n$ que pueden ser objetos arbitrarios. Medimos sus similitudes por pares $s_{ij} = s(x_i, x_j)$ mediante alguna función de similitud que sea simétrica y no negativa, y denotamos la matriz de similitud correspondiente como $S = (s_{ij})_{i,j=1,...,n}$.

Entrada: Matriz de similitud $S \in \mathbb{R}^{n \times n}$, número k de clústeres a construir.

- Construir un grafo de similitud. Sea **A** su matriz de adyacencia.
- Calcular el laplaciano **L** .
- Calcular los primeros k autovectores $u_1, ..., u_k$ de **L**.
- Formar la matriz $U \in \mathbb{R}^{n \times k}$ que contiene los vectores $u_1, ..., u_k$ como columnas.
- Para $i = 1, ..., n$, sea $y_i \in \mathbb{R}^k$ el vector correspondiente a la i-ésima fila de U.
- Agrupar los puntos $(y_i)_{i=1,...,n}$ en $\mathbb{R}^k$ clústeres utilizando el algoritmo k-means, obteniendo los clústeres $C_1, ..., C_k$.

Salida: Clústeres $A_1, ..., A_k$ donde $A_i = \{j |\ y_j \in C_i\}$.

Con el fin de ejemplificar el funcionamiento del algoritmo de agrupamiento espectral de manera clara y accesible, se presenta a continuación un ejemplo simplificado construido manualmente. Debido a que calcular todas las etapas del algoritmo —como la construcción de la matriz laplaciana, el cálculo de sus autovalores y autovectores, y la posterior transformación espectral— puede resultar algebraicamente extenso, se optó por comenzar desde una matriz de coordenadas reducida, que simula el resultado de las etapas previas del método.

Esta matriz, que puede interpretarse como una matriz $U \in R^{n \times k}$ compuesta por los k autovectores correspondientes a los menores autovalores de la matriz laplaciana, sirve como base para aplicar la última fase del algoritmo: el agrupamiento mediante k-means. Las filas de dicha matriz representan las muestras proyectadas en el espacio espectral, y sobre ellas se ejecuta el algoritmo de agrupamiento para particionar los nodos en k comunidades.

Este enfoque permite ilustrar de forma concreta y paso a paso cómo actúa el agrupamiento espectral a partir de su representación reducida, y al mismo tiempo mostrar el funcionamiento interno del algoritmo de k-means en un contexto más manejable.

Ejemplo 2.5.2. Sea la matriz de coordenadas

$$\begin{pmatrix} \text{Nodo} & \text{V1} & \text{V2} \\ N1 & 0,5 & 0,3 \\ N2 & 0,4 & 0,4 \\ N3 & 0,1 & 0,6 \\ N4 & -0,4 & -0,5 \\ N5 & -0,5 & -0,6 \\ N6 & 0,3 & 0,4 \\ N7 & 0,4 & 0,2 \\ N8 & -0,2 & -0,3 \\ N9 & -0,6 & -0,4 \end{pmatrix}$$

Elegimos $k = 2$ clusters.

Paso 1: Inicializar centroides.
Tomamos 2 puntos elegidos al azar como centroides iniciales.
$C1 = N1 = (0,5; 0,3)$
$C2 = N4 = (-0,4; -0,5)$

Paso 2: Calcular distancia a los centroides.

$$dist(x, y) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Calculamos la distancia de $N2$ a los centroides.

$$\begin{aligned} dist(N2, C1) &= \sqrt{(0,4 - 0,5)^2 + (0,4 - 0,3)^2} = \sqrt{0,01 + 0,01} = \sqrt{0,02} \cong 0,141 \\ dist(N2, C2) &= \sqrt{(0,4 + 0,4)^2 + (0,4 + 0,5)^2} = \sqrt{0,64 + 0,81} = \sqrt{1,45} \cong 1,204 \end{aligned}$$

Tenemos que $N2$ está más cerca de $C1$. El resto de los resultados nos dan:

$$\begin{pmatrix} \text{Punto} & \text{Distancia a } C1 & \text{Distancia a } C2 & \text{Más cerca de} \\ N1 & 0 & 1,204 & C1 \\ N2 & 0,141 & 1,204 & C1 \\ N3 & 0,5 & 1,208 & C1 \\ N4 & 1,204 & 0 & C2 \\ N5 & 1,345 & 0,141 & C2 \\ N6 & 0,223 & 1,140 & C1 \\ N7 & 0,141 & 1,063 & C1 \\ N8 & 0,922 & 0,282 & C2 \\ N9 & 1,303 & 0,223 & C2 \end{pmatrix}$$

Tenemos que los puntos $N1, N2, N3, N6$ y $N7$ están más cerca de $C1$, mientras que $N4, N5, N8$ y $N9$ lo están de $C2$.

Paso 3: Calculamos los nuevos centroides como la media de los puntos asignados a cada uno.
Para $C1$

$$\begin{aligned} \frac{0,5 + 0,4 + 0,1 + 0,3 + 0,4}{5} &= \frac{1,7}{5} = 0,34 \\ \frac{0,3 + 0,4 + 0,6 + 0,4 + 0,2}{5} &= \frac{1,9}{5} = 0,38 \end{aligned}$$

Por lo tanto $C1 = (0,34; 0,38)$.

Para $C2$

$$\begin{aligned}\frac{-0,4-0,5-0,2-0,6}{4} &= \frac{-1,7}{4} = -0,42\\ \frac{-0,5-0,6-0,3-0,4}{4} &= \frac{-1,8}{4} = -0,45\end{aligned}$$

Por lo que $C2 = (-0,42; -0,45)$.

De esta forma los nuevos centroides son $C1 = (0,34; 0,38)$ y $C2 = (-0,42; -0,45)$. Ahora repetiremos el paso 2 con los nuevos centroides.

$$\begin{pmatrix}\text{Punto} & \text{Distancia a } C1 & \text{Distancia a } C2 & \text{Más cerca de}\\ N1 & 0,178 & 1,190 & C1\\ N2 & 0,063 & 1,184 & C1\\ N3 & 0,325 & 1,173 & C1\\ N4 & 1,149 & 0,055 & C2\\ N5 & 1,290 & 0,167 & C2\\ N6 & 0,044 & 1,117 & C1\\ N7 & 0,189 & 1,050 & C1\\ N8 & 0,868 & 0,270 & C2\\ N9 & 1,221 & 0,182 & C2\end{pmatrix}$$

Como los puntos asignados a cada clúster son los mismos entonces debe converger en la segunda iteración, así que tras todo el proceso obtuvimos dos comunidades con los siguientes puntos: Comunidad 1: $N1, N2, N3, N6, N7$ y Comunidad 2: $N4, N5, N8, N9$.

3 Metodología

3.1 Fuentes de datos

La presente investigación tiene como objetivo aplicar y comparar distintas técnicas de detección de comunidades en grafos construidos a partir de datos relacionales. Para ello, se trabajó con una base de datos real del sistema de ciencia y tecnología argentino, así como con una red simulada de colaboración académica. El análisis se estructura en torno a tres enfoques: el coloreo de grafos como paso preliminar para organizar la información, el agrupamiento espectral como técnica basada en álgebra lineal, y el algoritmo de Girvan-Newman como método clásico centrado en la estructura de la red. A continuación, se detalla el procedimiento adoptado, los programas utilizados y las decisiones metodológicas tomadas en cada etapa del trabajo.

La base de datos utilizada corresponde al personal de ciencia y tecnología del año 2018, provista por el Sistema de Información de Ciencia y Tecnología Argentino (SICYTAR). Esta base fue obtenida del portal oficial datos.gob.ar.

La base contiene información detallada sobre todos los empleados vinculados al sector de ciencia y tecnología, organizada en diversas categorías o atributos. Entre estos atributos se encuentran: sexo, edad, grado académico, disciplina, tipo de personal, categoría del CONICET, dedicación horaria semanal, clase de cargo docente y tipo de condición docente. Cada categoría cuenta con identificadores numéricos (IDs) que permiten clasificar rápidamente la información, considerando el volumen total de personal (aproximadamente 68.000 registros). Por ejemplo, en la categoría "sexo", el ID 1 representa "femenino" y el ID 2 "masculino", y así sucesivamente en el resto de los atributos. El código para la construcción de ambas redes fue escrito en el software R utilizando el paquete igraph, y dada la cantidad de nodos involucrados, se utilizó el software Gephi para la visualización del grafo.

3.2 Análisis de los datos

Para aplicar la técnica de coloreo de grafos, se buscó emplearla como herramienta para organizar y visualizar mejor la información. Con este fin, los atributos fueron agrupados en cuatro conjuntos:

- Grupo 1: sexo, edad.
- Grupo 2: grado académico, disciplina.
- Grupo 3: tipo de personal, categoría del CONICET.
- Grupo 4: dedicación horaria semanal, clase de cargo docente, tipo de condición docente.

A partir de esta clasificación, se construyó un grafo bipartito. Por un lado, se ubicaron los nodos que representan a las personas, y por el otro, los nodos correspondientes a los atributos, divididos en los cuatro grupos previamente definidos. Cada nodo de persona fue conectado mediante cuatro aristas a un nodo por grupo, según los IDs correspondientes. A cada grupo se le asignó un color distinto, utilizando así la lógica del coloreo de grafos, que establece que a nodos adyacentes se les deben asignar colores diferentes.

Con el grafo bipartito generado, se aplicó la técnica de agrupamiento espectral para identificar comunidades. Dado que el grafo modela la similitud a través de relaciones compartidas —es decir, si dos personas comparten atributos, estarán estructuralmente próximas en el grafo—, la matriz laplaciana captura dicha estructura, y sus autovectores reflejan esta "proximidad estructural". Por tanto, resulta coherente aplicar un enfoque espectral, ya que este explora la estructura global del grafo a partir de su espectro. No obstante, debido a la gran cantidad de nodos, el entorno de R no puede procesar una matriz laplaciana de tal tamaño. Por ello, se extrajo una muestra de 200 nodos para la aplicación del algoritmo. En R se calcularon la matriz laplaciana y sus autovectores. Luego, utilizando la función kmeans() y seleccionando k = 4 como número de comunidades —valor determinado como óptimo a partir del método del codo, aplicado mediante la función fviz_nbclust() del paquete factoextra— se

ejecutó el algoritmo de k-means. La visualización de las comunidades resultantes también se realizó en Gephi.

Finalmente, para aplicar el algoritmo de Girvan-Newman, se diseñó una base de datos simulada compuesta por 40 personas, dado que la base original, al estar representada mediante un grafo bipartito sin conexiones entre nodos del mismo tipo, no permitía aplicar dicho algoritmo. En esta nueva base, los nodos representan investigadores del sistema de ciencia y tecnología que trabajan en un centro multidisciplinario, y las aristas simbolizan vínculos de colaboración directa (como coautorías de artículos, participación conjunta en proyectos, entre otros). El grafo fue implementado en R, y se aplicó el algoritmo utilizando la función cluster_edge_betweenness(). Posteriormente, a través de la función tkplot() se visualizaron las comunidades identificadas, así como también se mostrará el resto de los grafos en la sección de resultados.

4 Resultados

4.1 Grafo coloreado

En la Figura 30 se visualiza el grafo bipartito construido a partir de la base de datos del personal de ciencia y tecnología del año 2018. Los nodos celestes representan a las personas, mientras que los demás colores corresponden a los cuatro grupos de atributos definidos previamente: púrpura para el grupo 1 (sexo y edad), verde para el grupo 2 (grado académico y disciplina), rojo para el grupo 3 (tipo de personal y categoría del CONICET), y naranja para el grupo 4 (dedicación horaria, clase y condición docente). La construcción en forma bipartita permite vincular a cada persona con los atributos que cumple, lo que facilita posteriormente asignarlas a grupos según determinados requisitos y avanzar en el análisis de comunidades. Cabe señalar que este grafo se elaboró a partir de aproximadamente 68.000 registros, y que su construcción tiene como objetivo organizar la información mediante el coloreo de grafos, utilizado aquí como punto de partida para el agrupamiento y posterior detección de comunidades.

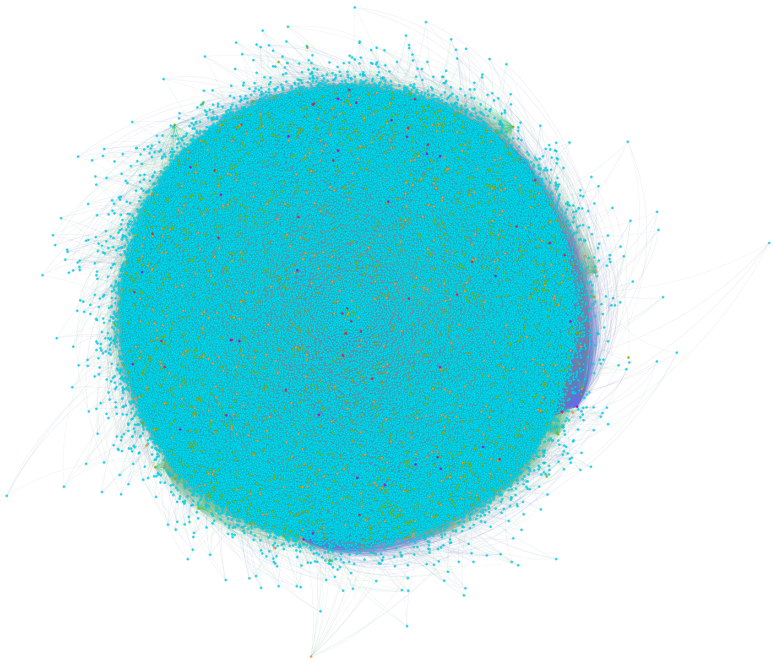

Figura 34: Grafo bipartito de la base de datos

En la Figura 31 se observa un sector de la red con mayor detalle, mientras que en la Figura 32 se muestra un nodo correspondiente a una persona junto con sus cuatro aristas.

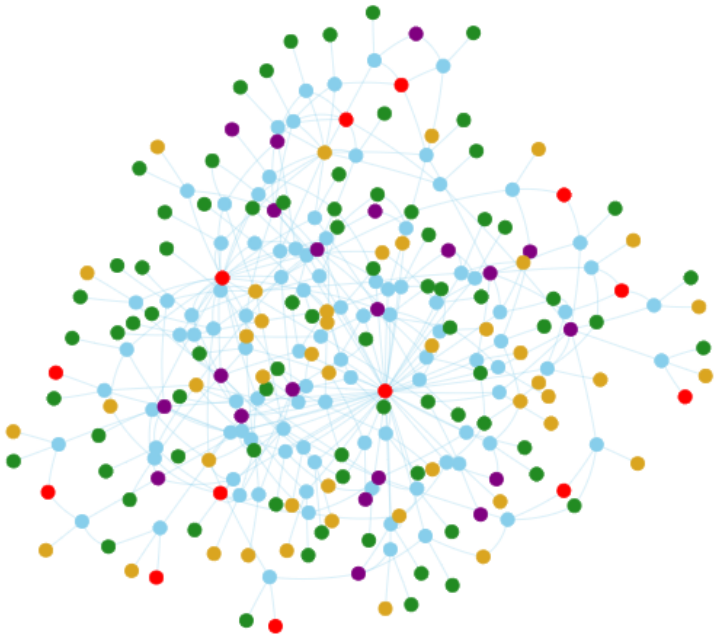

Figura 35: Sector de la red ampliado

Cada uno de los nodos conectados tiene el color asociado a su grupo y una etiqueta con sus identificadores (ID), que permiten acceder a la información asociada de esa persona. Por ejemplo, el nodo púrpura (grupo 1 o A) posee la descripción " A_1_9", lo que indica que la persona número 8872 es de sexo femenino (ID 1 del atributo "sexo") y tiene una edad comprendida entre 58 y 62 años (ID 9 de la categoría "edad"). De manera análoga, los demás nodos y sus aristas se interpretan siguiendo este mismo criterio.

Figura 36: Nodo de persona con sus cuatro aristas

Este grafo fue generado utilizando el paquete igraph de R, y visualizado mediante Gephi, aprovechando su capacidad para manejar grandes volúmenes de nodos.

La aplicación de la teoría del coloreo de grafos permitió organizar visualmente la red, asignando un color distinto a cada grupo de atributos. Esto facilitó no solo una lectura más clara de la estructura general, sino también la identificación visual de la diversidad y distribución de perfiles dentro del sistema científico. El objetivo de este esquema de coloreo fue doble: por un lado, garantizar que cada nodo persona estuviera conectado a exactamente un nodo de cada grupo, representando así la completitud de su perfil individual; por otro lado, preparar la estructura para una posterior aplicación de técnicas de detección de comunidades. En ese sentido, esta fase funcionó como un paso organizativo fundamental para el análisis espectral realizado sobre una muestra reducida del grafo, como se detallará en la sección siguiente.

4.2 Detección de comunidades con agrupamiento espectral

Para proceder con la detección de comunidades en el grafo real, se optó por una combinación de análisis espectral y el algoritmo de k-means. En primer lugar, se aplicó el método del codo para determinar el número óptimo de clústeres a utilizar. Como se observa en la Figura 33, el punto de inflexión en la curva se encuentra en k = 4, lo que sugiere que cuatro es la cantidad adecuada de comunidades para representar la estructura subyacente del grafo.

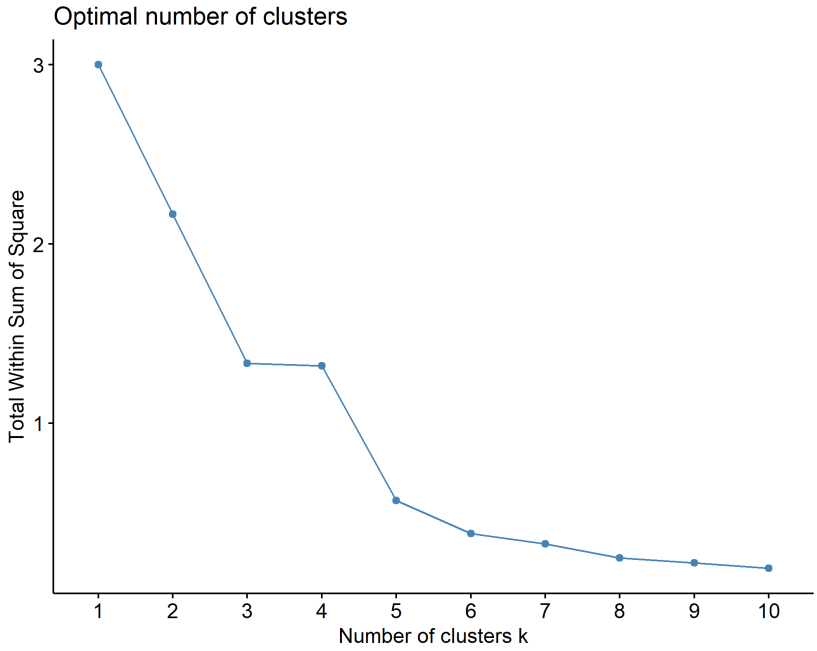

Figura 37: Método del codo

A continuación, se realizó el agrupamiento espectral, extrayendo los vectores propios asociados a los menores autovalores del Laplaciano del grafo, para luego aplicar sobre ellos k-means con k=4. El resultado puede verse en la Figura 34, donde cada nodo ha sido coloreado según la comunidad a la que fue asignado: rojo, verde, celeste y violeta.

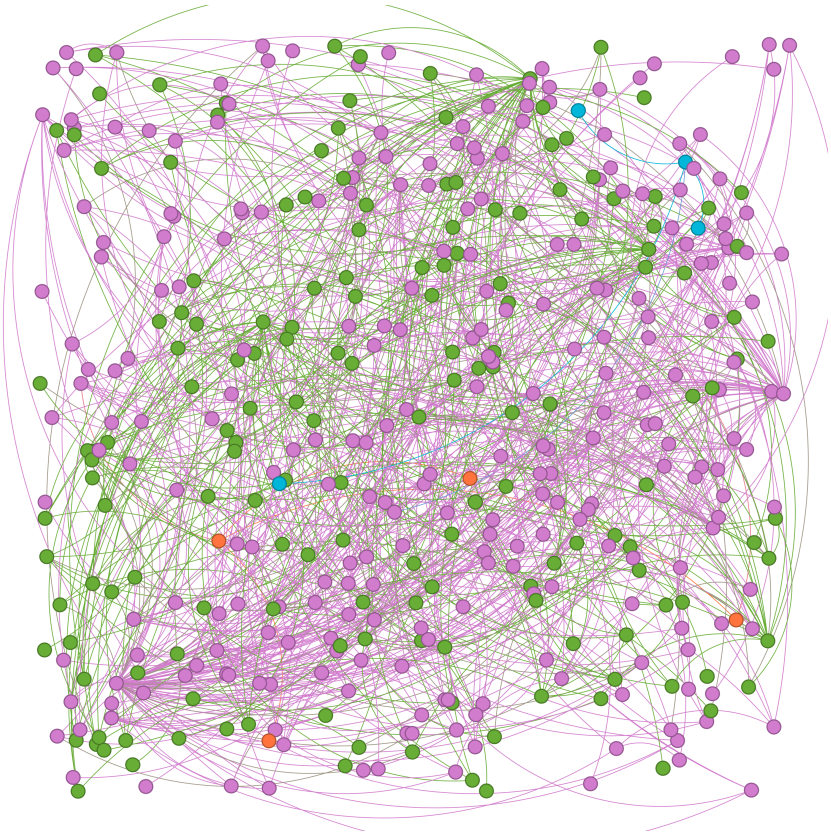

Figura 38: Red con comunidades identificadas tras aplicar agrupamiento espectral

Se advierte una distribución desigual en el tamaño de las comunidades: las comunidades 2 (verde) y 4 (violeta) concentran la mayoría de los nodos, mientras que las comunidades 1 (rojo) y 3 (celeste) contienen solo cuatro nodos cada una. Esta asimetría sugiere que algunas regiones del grafo presentan mayor densidad y cohesión interna, mientras que otras podrían estar formadas por nodos periféricos o menos conectados entre sí, posiblemente agrupados por similitud espectral más que por conectividad directa.

En este marco, la distribución desigual de las comunidades puede interpretarse en el contexto de la composición del sistema de ciencia y tecnología: las comunidades más numerosas probablemente correspondan a perfiles mayoritarios de investigadores que comparten atributos frecuentes-por ejemplo, combinaciones habituales de disciplina, grado académico y cargo docente-, mientras que las comunidades más pequeñas reflejan perfiles minoritarios o periféricos, definidos por combinaciones menos comunes. Esto muestra que, más allá de la estructura matemática del grafo, el análisis también permite visibilizar concentraciones y particularidades en la distribución de perfiles dentro del sistema.

4.3 Grafo simulado y comunidades identificadas con Girvan-Newman

Con la nueva base de datos simulada, compuesta por 40 investigadores en un centro multidisciplinario y conectados a través de vínculos de colaboración académica (como coautorías o proyectos compartidos), se construyó el siguiente grafo de 40 nodos (Figura 35).

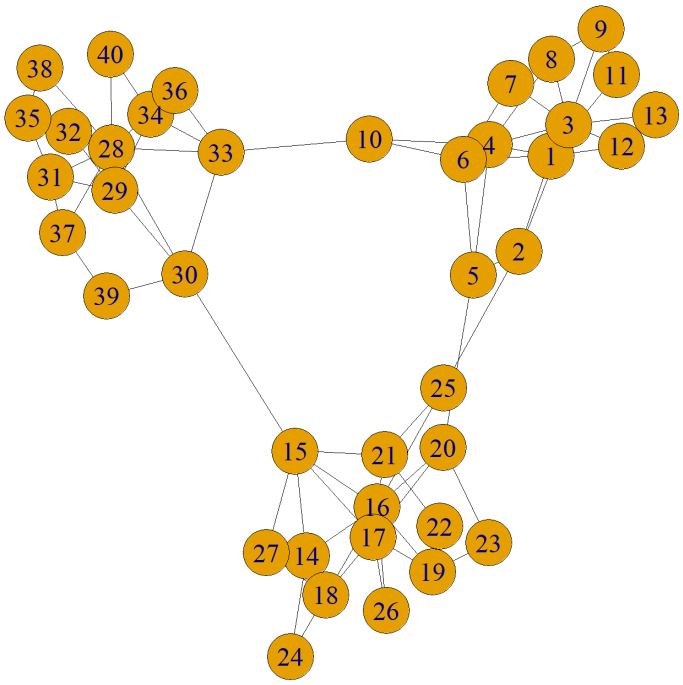

Figura 39: Grafo de 40 nodos antes de aplicar el algoritmo de Girvan-Newman

A simple vista, puede observarse que la estructura del grafo presenta una forma triangular, con tres agrupamientos locales que ya sugieren una organización comunitaria. Aunque aún no se han identificado explícitamente las comunidades, la disposición espacial de los nodos comienza a mostrar cierta densidad interna y conexiones más escasas entre agrupamientos.

Tras aplicar el algoritmo de Girvan-Newman, se obtiene una segmentación clara del grafo en tres comunidades (Figura 36). La comunidad 1, de color naranja, incluye 13 nodos que se concentran en la parte inferior izquierda del grafo. La comunidad 2, de color azul, también con 14 nodos, se ubica mayormente en la parte superior. Por último, la comunidad 3, de color verde, con 13 nodos, se sitúa en el extremo derecho del grafo.

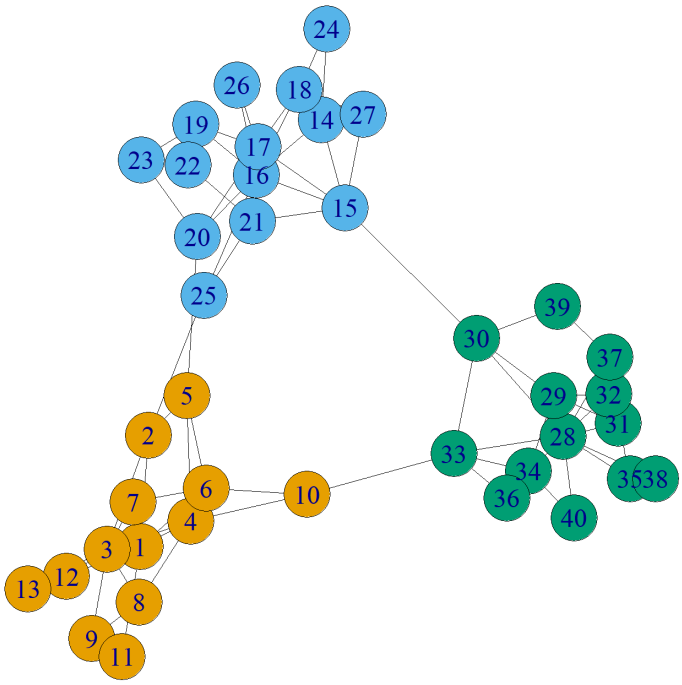

Figura 40: Grafo tras aplicar el algoritmo de Girvan-Newman

Lo interesante del resultado es que la segmentación coincide con las zonas de mayor densidad de conexiones locales observadas en la visualización previa. Cada comunidad forma un subgrafo cohesivo, con varias conexiones internas y relativamente pocas conexiones con los otros grupos. Además, los nodos intercomunitarios (por ejemplo, el nodo 10 y el nodo 33) se ubican estratégicamente como puntos de enlace o puentes entre comunidades, lo que los vuelve potenciales nodos clave en la estructura global del grafo.

Si se interpreta el grafo en el contexto de redes de colaboración, puede pensarse que cada comunidad representa un conjunto de investigadores con afinidades temáticas o de trabajo. Por ejemplo:

- La comunidad naranja podría representar un núcleo de investigadores en ciencias exactas.
- La azul, un grupo enfocado en disciplinas aplicadas y tecnológicas.
- La verde, un conjunto de investigadores orientados a las ciencias sociales o la extensión universitaria.

Esta segmentación refuerza la idea de que la estructura de la red no es aleatoria, sino que refleja patrones de afinidad, colaboración o disciplina, aún en una simulación. Así, el análisis no solo permite validar la técnica de detección de comunidades, sino que también aporta una herramienta conceptual para interpretar las dinámicas internas de grupos de trabajo académico. La representación visual del grafo antes y después del análisis comunitario ayuda a hacer tangible esta organización latente.

5 Conclusión

El presente trabajo logró cumplir con los objetivos propuestos en relación con la exploración y aplicación de diversas estrategias de análisis estructural de redes complejas. A partir del uso de datos relacionales reales (SICYTAR, 2018) y de redes simuladas, se implementaron y compararon tres enfoques distintos, cada uno con aportes específicos al estudio de la estructura interna de las redes.

En primer lugar, tal y como se había propuesto, el coloreo de grafos permitió organizar las redes para la identificación de estructuras y facilitar la interpretación visual de las mismas al ordenarlas de manera preliminar, ofreciendo así una partición inicial que favoreció la interpretación visual y la comprensión de las relaciones entre nodos. Aunque no constituye un método formal de detección de comunidades, se comprobó que su aplicación facilita la identificación de bloques independientes y la preparación del grafo para análisis posteriores, mostrando además su utilidad práctica en la organización de tareas y en contextos colaborativos.

Posteriormente, el agrupamiento espectral evidenció el poder de las herramientas algebraicas en la detección de patrones latentes, cumpliendo así el segundo objetivo de detectar comunidades y patrones latentes con este método, como herramienta algebraica basada en propiedades de matrices asociadas a los grafos. Mediante el análisis de la matriz laplaciana y sus autovalores, se logró segmentar la red en comunidades con una base matemática sólida. Este enfoque demostró ser especialmente valioso para resaltar estructuras internas que no resultan evidentes a simple vista, confirmando la relevancia del vínculo entre álgebra lineal y teoría de grafos en el estudio de redes complejas.

Finalmente, la aplicación del algoritmo de Girvan–Newman permitió identificar comunidades a partir de la eliminación de aristas con alta intermediación, lo cual era el tercer objetivo. Esto brindó una perspectiva complementaria al enfoque espectral. Los resultados obtenidos con este método sobre la red simulada de 40 nodos ilustraron cómo la eliminación progresiva de conexiones críticas revela divisiones naturales dentro de la estructura, aportando un contraste interesante respecto de los métodos algebraicos.

En conjunto, la integración de estas tres técnicas mostró que es posible articular diferentes enfoques en un marco analítico común para estudiar redes complejas, alcanzando de esta manera el objetivo general de la tesis que era explorar y aplicar distintas estrategias de análisis estructural en las redes, haciendo énfasis en estas tres perspectivas. Cada método aportó una lectura distinta: el coloreo como estrategia organizativa, el agrupamiento espectral como herramienta algebraica de segmentación y Girvan–Newman como algoritmo clásico de detección. La combinación de resultados permitió no solo corroborar patrones internos en las redes analizadas, sino también evidenciar la complementariedad de los métodos.

Así, la tesis no solo alcanzó sus objetivos, sino que también dejó en evidencia que la aplicación conjunta de herramientas de teoría de grafos y álgebra lineal ofrece una vía prometedora para optimizar el análisis de datos relacionales en distintos ámbitos, desde instituciones científicas hasta entornos educativos. Como líneas futuras, resultaría de interés ampliar la escala de las redes analizadas y explorar la integración de estas técnicas con otros métodos, lo que podría enriquecer aún más la comprensión y división de redes complejas.

A partir de los resultados obtenidos, se abren diversas proyecciones a futuro que podrían fortalecer y ampliar el alcance de la investigación. En primer lugar, sería valioso replicar el análisis sobre redes de mayor escala y diversidad estructural, incorporando datos provenientes de distintos años, áreas disciplinares o instituciones. Esto permitiría evaluar la solidez de los métodos aplicados y analizar cómo varían las comunidades detectadas en función de la naturaleza de las relaciones y del tamaño de la red.

En segundo lugar, se propone integrar las técnicas de teoría de grafos y álgebra lineal con métodos de aprendizaje automático y minería de datos, con el fin de automatizar la detección de patrones y profundizar en la interpretación de los vínculos. Este enfoque híbrido podría facilitar la identificación de comunidades latentes, mejorar la clasificación de nodos según métricas combinadas y abrir el camino hacia modelos predictivos en el análisis de redes complejas.

Además, una posible proyección a futuro diferente sería incorporar otros métodos de detección de

comunidades, como el algoritmo de Louvain, que permite detectar comunidades en redes de gran escala. Esto permitiría comparar sus resultados con los obtenidos por Girvan–Newman y el agrupamiento espectral, y evaluar qué enfoque ofrece una mejor división estructural según el tipo y tamaño de la red

Finalmente, otra línea de investigación a futuro prometedora podría consistir en trasladar los resultados teóricos y computacionales a contextos educativos y de gestión, desarrollando herramientas didácticas o aplicaciones prácticas basadas en grafos que promuevan el pensamiento crítico, la organización colaborativa y la toma de decisiones informada. De esta manera, la integración entre teoría de grafos, álgebra lineal y análisis de datos podría consolidarse no solo como una estrategia de investigación, sino también como una herramienta formativa y de optimización en distintos ámbitos del conocimiento.

6 Bibliografía

Aggarwal, C. C., & Wang, H. (2010). *Managing and mining graph data*. Springer.https://doi.org/10.1007/978-1-4419-6045-0

Appel, K., & Haken, W. (1977). Every planar map is four colorable. Part I: Discharging. *Illinois Journal of Mathematics*, 21(3), 429–490.

Bapat, R. B. (2010). *Graphs and matrices*. Springer. https://doi.org/10.1007/978-1-84882-981-7

Braicovich, T., Caro, P., Cerda, V., Oropeza, M., Osio, E. & Reyes, C. (2009). *Introducción a la Teoría de Grafos*. educo.

Carrasco-Pilco, L.F., Burgos-Cevallos, V.E., Jurado-Liberona, G. & Nymoen-Bonilla, E.N. (2021). Coloración de Grafos y su aplicación a la Geografía. *Polo del Conocimiento*, 6 (9), 1519-1544. https://polodelconocimiento.com/ojs/index.php/es/article/view/3125

Chaitin, G. J. (1982). Register allocation & spilling via graph coloring. *ACM SIGPLAN Notices*, 17(6), 98–101. https://doi.org/10.1145/872726.806984

Chung, F. R. K. (1997). *Spectral graph theory* (Vol. 92). American Mathematical Society.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3–5), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12), 7821–7826.https://doi.org/10.1073/pnas.122653799

Holme, P. (2005). Network reachability of real-world contact sequences. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 71(4), 046119. https://doi.org/10.48550/arXiv.cond-mat/0410313

Horn, R. A., & Johnson, C. R. (2013). *Matrix analysis* (2nd ed.). Cambridge University Press.

Iacobucci, D. (2013). Grafos y matrices. En S. Wasserman & K. Faust, *Social network analysis: Methods and applications* (pp. 92–166). Cambridge University Press.

Jerónimo, G., Sabía, J., & Tesauri, S. (2008). *Álgebra lineal*. EUDEBA.

Kannan, M., Nivetha, P., Sankar, K., & Gurjar, J. (2024). Graph coloring techniques in scheduling and resource allocation. *Journal of Nonlinear Analysis and Optimization*, 15(2), 70–77.

Kirman, A. (2008). Graph theory. In *The New Palgrave Dictionary of Economics*. Palgrave Macmillan. https://doi.org/10.1057/978-1-349-95121-5_1232-2

Kumar, J. S., Archana, B., Muralidharan, K., & Senthil Kumar, V. (2025). Graph theory: Modelling and analyzing complex system. *Metallurgical and Materials Engineering*, 31(3), 70–77. https://doi.org/10.63278/1320

Lewis, R.M.R. (2016). *A Guide to Graph Colouring: Algorithms and Applications* (1.ª ed.). Springer.

Li, J., Lai, S., Shuai, Z., Tan, Y., Jia, Y., Yu, M., ... & Lu, Y. (2024). A comprehensive review of community detection in graphs. *Neurocomputing*, 600, 128169.https://doi.org/10.48550/arXiv.2309.11798

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Newman, M. (2010). *Networks: An introduction*. Oxford University Press.

Palacios, F., & Caro, P. (2025). Asignación inteligente de tareas con grafos: eficiencia en la administración de recursos humanos. *Cuadernos De Investigación Serie Administración*, (7), 87–97. https://revele.uncoma.edu.ar/index.php/administracion/article/view/6903

Pérez, J. (2022). *Coexistencia de matemática y salud: Grafos como modelizadores e indicadores de redes* [Tesis de licenciatura, Universidad Nacional del Comahue]. Repositorio Digital Institucional UNCo. http://rdi.uncoma.edu.ar/handle/uncomaid/17867

Poole, D. (2010). *Linear Algebra: A Modern Introduction* (3^a ed.). Cengage Learning.

Strang, G. (2009). *Introduction to linear algebra* (4th ed.). Wellesley-Cambridge Press.

Šumak, B., & Pušnik, M. (2023). Analysis of the shortest path method application in social networks. In *Information Modelling and Knowledge Bases XXXIV* (Vol. 364, pp. 169–182). Frontiers in Artificial Intelligence and Applications. https://doi.org/10.3233/FAIA220500

Torres, P. (s.f.). *Capítulo 3: Coloreo de Grafos*. Asignatura: Tópicos Avanzados en Teoría de Grafos. Universidad Nacional de Rosario.

Trujillo Oval, Z. (2021). *Coloración de Grafos* [Trabajo de Fin de Grado, Universidad de La Laguna]. http://riull.ull.es/xmlui/handle/915/24111

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4), 395–416. https://doi.org/10.1007/s11222-007-9033-z