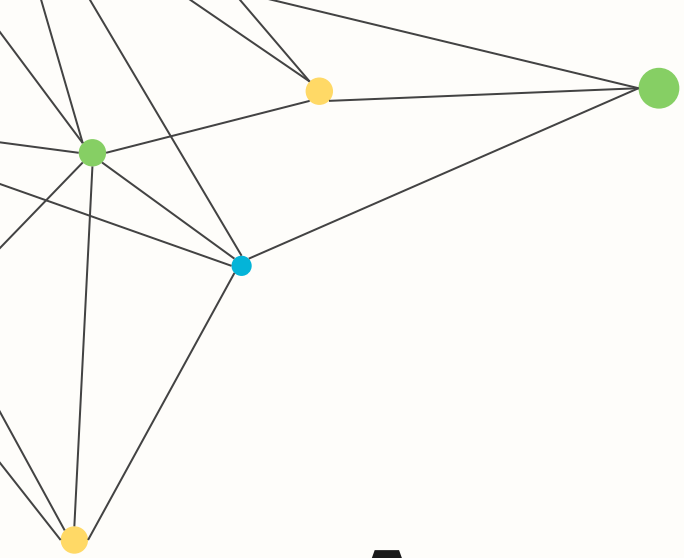
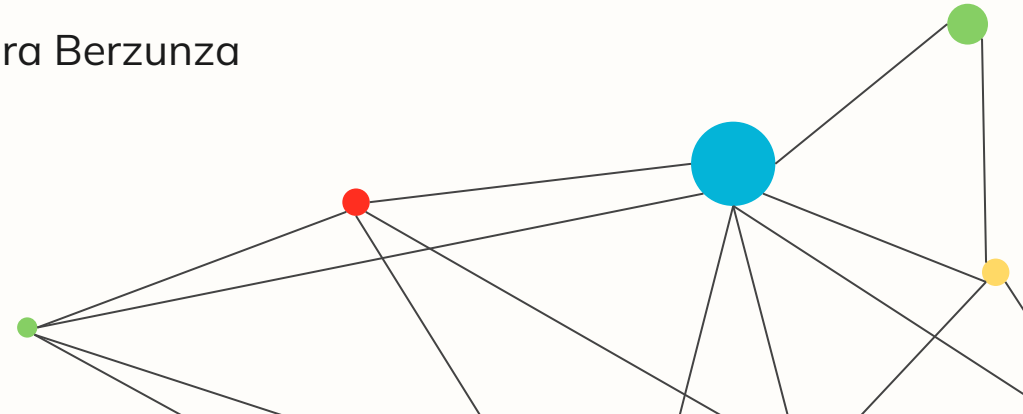


Agrupamiento en grafos

Alondra Berzunza



Agrupamiento de Grafos

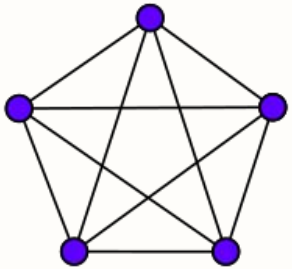
En teoría de grafos y análisis de redes sociales, el agrupamiento o agrupabilidad (en inglés, clustering) es una propiedad de un grafo o red social, que generaliza la noción de equilibrio estructural.

Los niveles de agrupabilidad se pueden cuantificar a través de coeficientes de agrupamiento. De estos conceptos han derivado en la actualidad diversos algoritmos de agrupamiento utilizados en minería de datos.

Coeficiente de Agrupamiento

En ciencia de redes, el coeficiente de agrupamiento de un vértice en un grafo cuantifica qué tanto está de agrupado (o interconectado) con sus vecinos.

Si el vértice está agrupado como un subgrafo completo, entonces su valor es máximo, mientras que un valor pequeño indica un vértice poco agrupado en la red.



El grafo completo K5. En un subgrafo como este, los vértices forman un clique de tamaño 5.

El tamaño de un clique es el número de vértices que contiene.

Se suele representar formalmente como Ci. En el análisis de redes sociales, en ocasiones a este coeficiente se le conoce también como transitividad.

Coeficiente de Agrupamiento

El coeficiente de agrupamiento es una medida de la cohesión en una red que cuantifica la tendencia de los nodos a agruparse, y se divide en coeficientes locales y globales.

Se calcula, en esencia, como la proporción de triángulos cerrados en una red, lo que indica qué tan conectados están los vecinos de un nodo entre sí. Un valor alto significa que los nodos de los que un nodo está conectado también tienden a estar conectados entre sí, como en una red social donde los amigos de una persona son amigos entre sí.

Coeficiente de Agrupamiento Local

- **Definición:** Mide la cohesión alrededor de un nodo específico.
- **Cálculo:** Se calcula como la proporción de aristas que existen entre los vecinos de un nodo, en relación con el número total de aristas posibles que podrían existir entre ellos.
- **Ejemplo:** En una red social, un coeficiente de agrupamiento local alto para una persona significa que sus amigos tienden a ser amigos entre s

Coeficiente de Agrupamiento Global

- **Definición:** Mide la tendencia de agrupamiento de toda la red.
- **Cálculo:** Se calcula como la suma normalizada de todos los coeficientes de agrupamiento locales.
- **Rango:** Varía entre 0 y 1. Un valor de 0 indica que la red está completamente desagrupada, mientras que un valor de 1 indica que cada nodo es parte de un triángulo cerrado.

Definición Formal

Un grafo signado es agrupable o tiene agrupamiento si sus nodos se pueden dividir en un número finito de subconjuntos (agrupamientos o clusters) tales que las arista positivas del grafo conectan a nodos en un mismo subconjunto, y las aristas negativas conectan a nodos en subconjuntos distintos.

Teorema 1. Un grafo signado no dirigido es agrupable si y sólo si no contiene ciclos con exactamente una arista negativa.

Teorema 2. Dado un grafo signado no dirigido completo, las siguientes aseveraciones son equivalentes:

- El grafo es agrupable.
- El grafo tiene un agrupamiento único.
- El grafo no tiene ningún ciclo con exactamente una arista negativa.
- El grafo no tiene ningún ciclo de longitud 3 con exactamente una arista negativa.

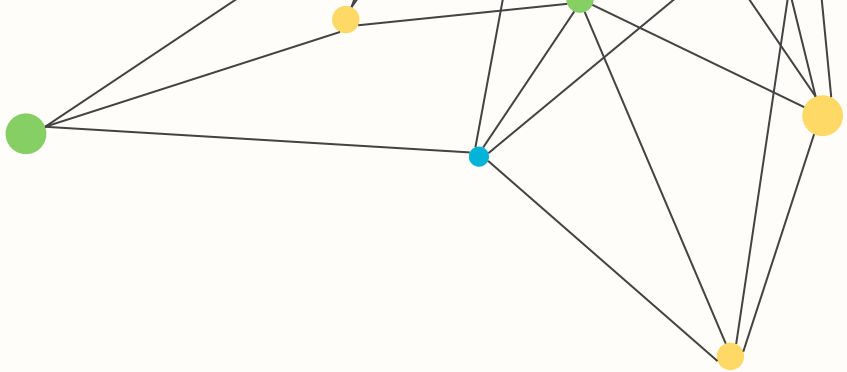
(si el grafo es dirigido, reemplazar «ciclo» por «semiciclo»)

Agrupamiento

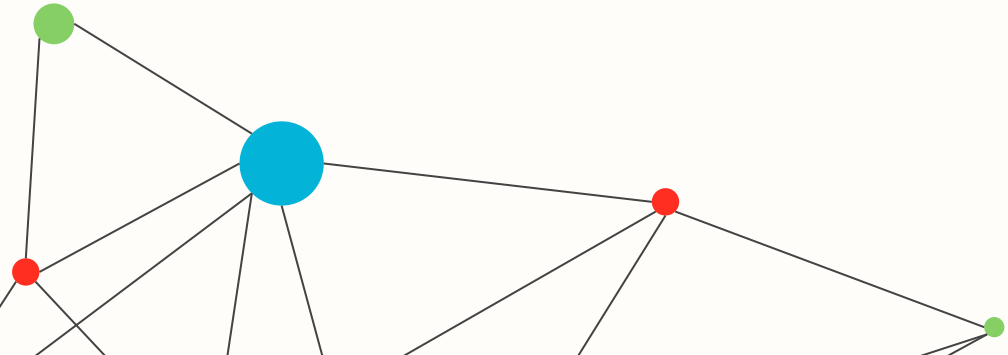
En el ámbito del análisis de datos, los algoritmos de agrupamiento de grafos se han convertido en herramientas poderosas para descubrir patrones, comunidades y estructuras dentro de redes complejas.

La agrupación de grafos es una rama del aprendizaje no supervisado dentro del aprendizaje automático que consiste en particionar los nodos de un grafo en grupos cohesivos (clústeres) según sus características comunes.

Como regla general, los nodos asignados al mismo clúster tienen más propiedades en común entre sí que con los nodos de clústeres diferentes. Según el método elegido, estas propiedades pueden ser valores predefinidos o calcularse implícitamente a partir de la estructura de conectividad del grafo.



Aplicaciones del Agrupamiento de Grafos



Aplicaciones de algoritmos de agrupamiento de grafos

Análisis de redes sociales: Estos algoritmos ayudan a identificar comunidades, individuos influyentes y líderes de opinión en las redes sociales. Por ello, se utilizan en marketing y también contribuyen a comprender la dinámica social .

Sistemas de recomendación: El uso de algoritmos de agrupamiento de grafos ayuda a agrupar objetos, personas o elementos similares según el comportamiento del usuario. Los sistemas de recomendación utilizan estos datos para ayudar a los usuarios a descubrir nuevas cuentas y recibir anuncios relevantes.

Aplicaciones de algoritmos de agrupamiento de grafos

Análisis de redes biológicas: Las redes de interacción proteína-proteína y las redes de expresión génica se representan computacionalmente como grafos, lo que significa que la agrupación de grafos es aplicable para identificar sus módulos funcionales. Esto permite a los investigadores comprender mejor las funciones e interacciones de sus componentes para profundizar en el estudio de los procesos biológicos.

Seguridad y detección de fraude: Los algoritmos de agrupamiento de grafos pueden detectar valores atípicos o anomalías dentro de las redes, lo que ayuda a identificar actividades fraudulentas, patrones de lavado de dinero o actores maliciosos. Con los datos que proporcionan, estos algoritmos ayudan a mantener la seguridad de la red y a prevenir delitos en el sector financiero y de seguros.

Es aplicable a cualquier problema que se pueda expresar en términos de redes y patrones contenidos en su estructura.

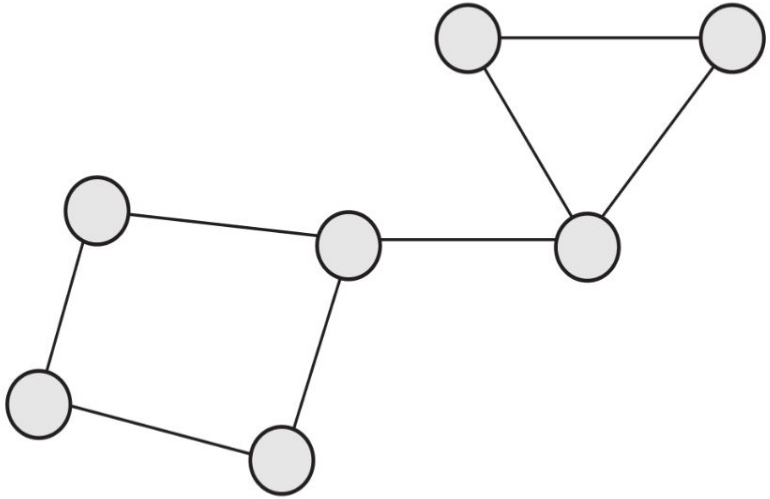
Agrupamiento

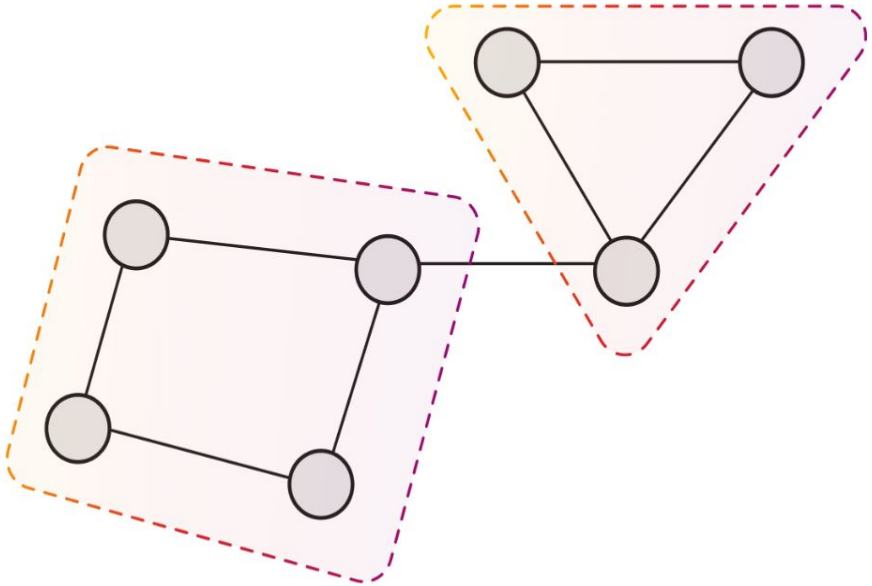
El proceso de agrupar elementos/entidades que, según una medida de similitud, parecen estar más próximos entre sí (que con respecto a los demás elementos) se denomina **análisis de conglomerados.**

La medida de similitud suele calcularse con base en criterios topológicos, como la estructura del grafo, la ubicación de los nodos u otras características, como las propiedades específicas de los elementos del grafo.

Los nodos que, según este valor de similitud, se consideran similares se agrupan en clústeres .

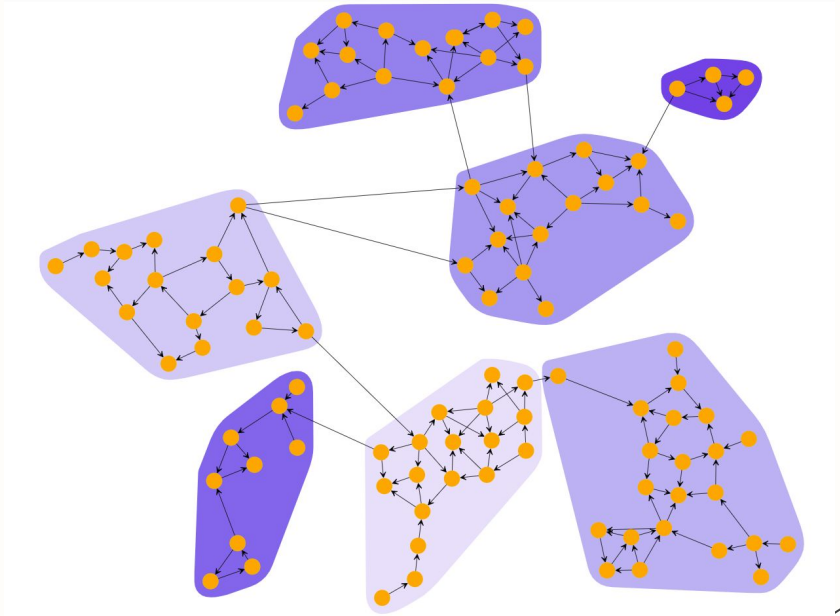
En otras palabras, cada clúster contiene elementos que comparten propiedades y características comunes. El conjunto de todos los clústeres constituye una agrupación .





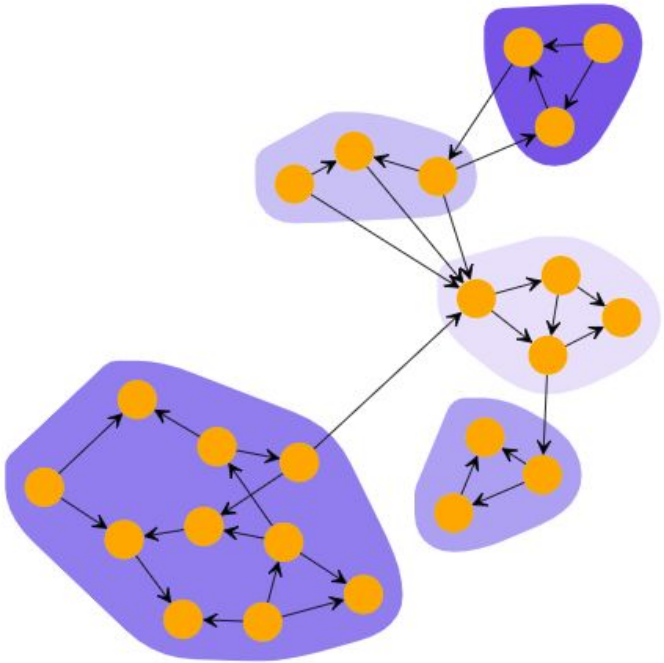
Agrupamiento de intermediación de bordes

La agrupación por intermediación de aristas detecta clústeres en una red de grafos eliminando progresivamente la arista con mayor centralidad de intermediación. La centralidad de intermediación mide la frecuencia con la que un nodo/arista se encuentra en la ruta más corta entre cada par de nodos del diagrama. El método se detiene cuando ya no quedan aristas que eliminar o si el algoritmo alcanza el número máximo de clústeres solicitado.



Agrupación de componentes biconectados

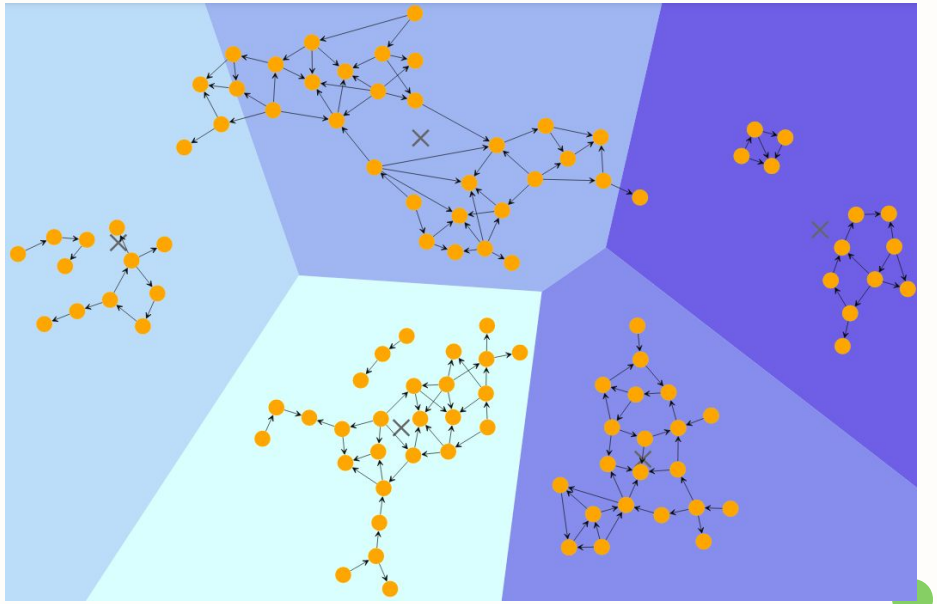
La agrupación de componentes biconectados detecta clústeres mediante el análisis de los componentes biconectados del grafo. Un componente biconectado es un componente conectado cuya propiedad es que la eliminación de cualquier nodo lo mantiene conectado. Los nodos de un mismo componente biconectado forman un clúster. Si un nodo pertenece a varios componentes biconectados, el algoritmo solo lo asigna a uno de estos clústeres.



Agrupamiento de k-medias

El algoritmo de agrupamiento k-medias divide el grafo en k grupos según la ubicación de los nodos, de modo que su distancia a la media (centroide) del grupo sea mínima. La distancia se define mediante diversas métricas.

También se puede implementar un diagrama de Voronoi. Consiste en dividir el plano que contiene n puntos (llamados sitios o generadores) en regiones poligonales convexas, de modo que cada región contenga exactamente uno de estos puntos, y cada punto del polígono esté más cerca del punto generador que de cualquier otro.



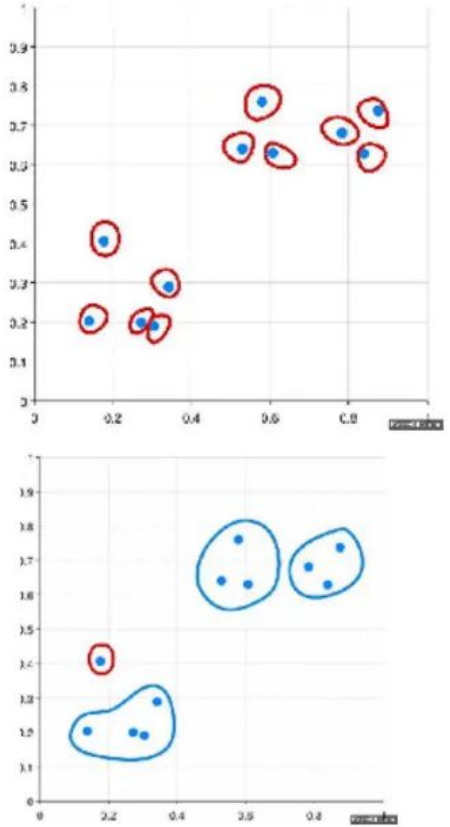
Agrupamiento jerárquico

La agrupación jerárquica agrupa los nodos en una estructura arborizada según su conectividad o similitud. A diferencia de los métodos de agrupación plana, proporciona una jerarquía anidada de clústeres, lo que la hace útil para aplicaciones que requieren agrupación multinivel.

Este método sigue dos enfoques principales:

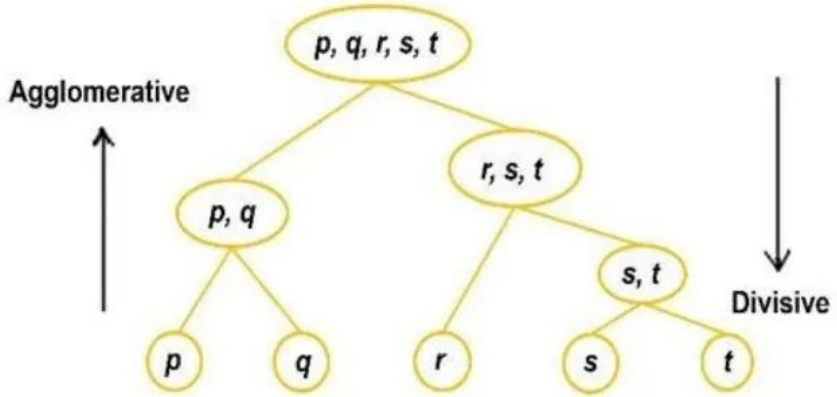
Aglomerativo

- Cada nodo comienza como su propio clúster.
- El algoritmo fusiona iterativamente los dos grupos más similares basándose en una medida de similitud (por ejemplo, ruta más corta, peso del borde o conectividad).
- Esta fusión continúa hasta que todos los nodos pertenecen a un solo clúster o se alcanza una cantidad predefinida de clústeres.
- Los ejemplos de medidas de similitud incluyen enlace simple (distancia mínima), enlace completo (distancia máxima) y enlace promedio.



Divisivo

- Todo el gráfico se trata inicialmente como un solo grupo.
- Se divide recursivamente en grupos más pequeños utilizando una técnica de partición de gráficos (por ejemplo, algoritmos de corte mínimo).
- El proceso se detiene cuando cada grupo cumple un criterio de detención, como un número máximo de grupos o una similitud mínima dentro del grupo.



Agrupamiento jerárquico

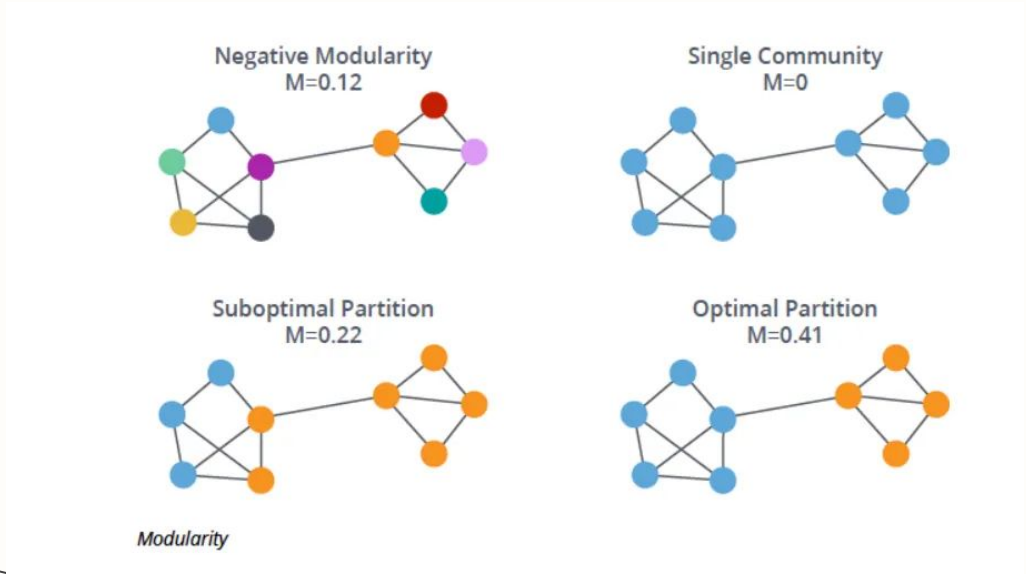
Los métodos modernos de agrupamiento jerárquico a gran escala utilizan:

- Matrices de similitud dispersas para reducir el uso de memoria.
- Técnicas de vecino más cercano aproximado (ANN) para evitar calcular todas las similitudes por pares.
- Implementaciones paralelas y distribuidas (por ejemplo, el método de agrupamiento de billones de bordes de Google) para gestionar conjuntos de datos masivos de manera eficiente.

Para escalar la agrupación jerárquica de gráficos de billones de aristas se requieren métodos avanzados como la esparsificación de gráficos, estrategias de vinculación escalables y marcos de computación distribuida (Google Blog, 2024).

Algoritmos basados en modularidad

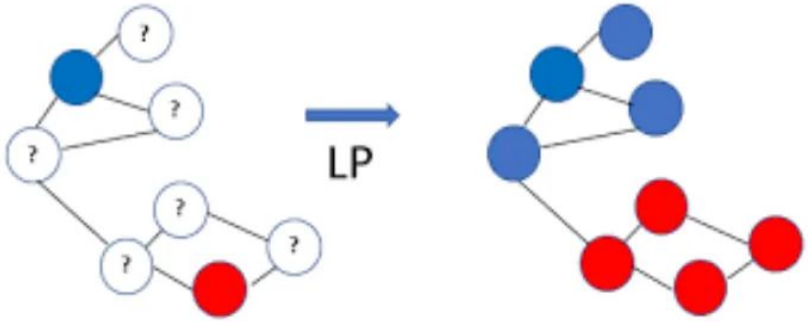
Los algoritmos basados en modularidad buscan optimizar la modularidad, una métrica que evalúa la calidad de las particiones de grafos. Por ejemplo, el algoritmo de Girvan-Newman identifica comunidades eliminando iterativamente las aristas con mayor centralidad de intermediación, lo que revela eficazmente la estructura subyacente del grafo.

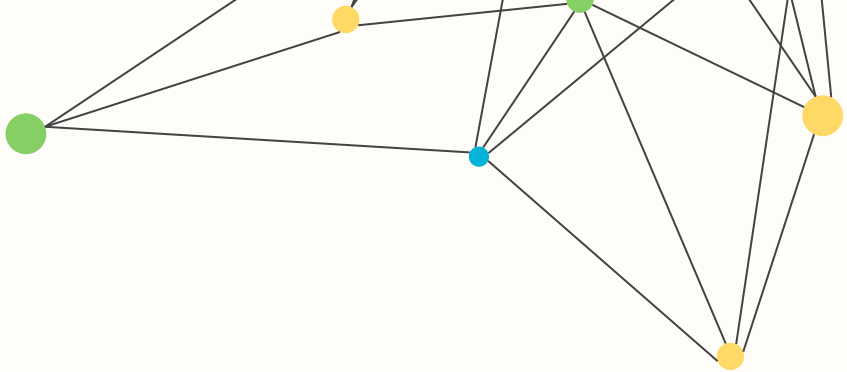


Propagación de etiquetas

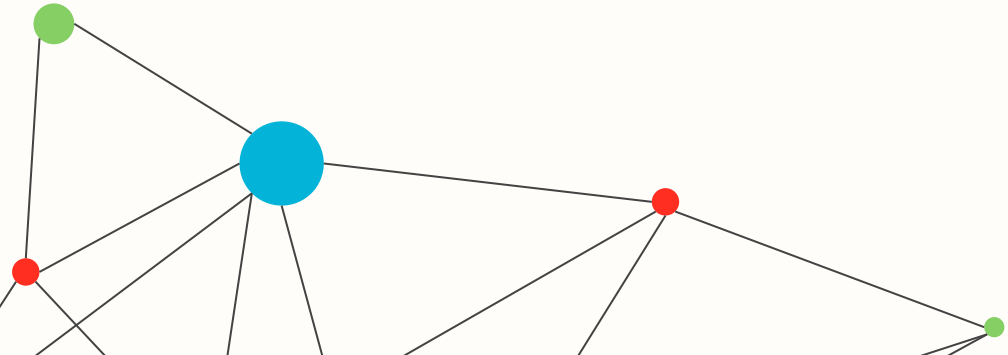
La propagación de etiquetas (LP) es un algoritmo iterativo simple, eficiente y escalable.

Cada nodo posee inicialmente una etiqueta única y, durante cada iteración, adopta la etiqueta más frecuente de sus vecinos. Este proceso continúa hasta que las etiquetas se estabilizan y no se producen más cambios. La propagación de etiquetas se utiliza ampliamente en grafos a gran escala debido a su velocidad y capacidad para gestionar redes extensas.





Métodos avanzados de agrupamiento de gráficos



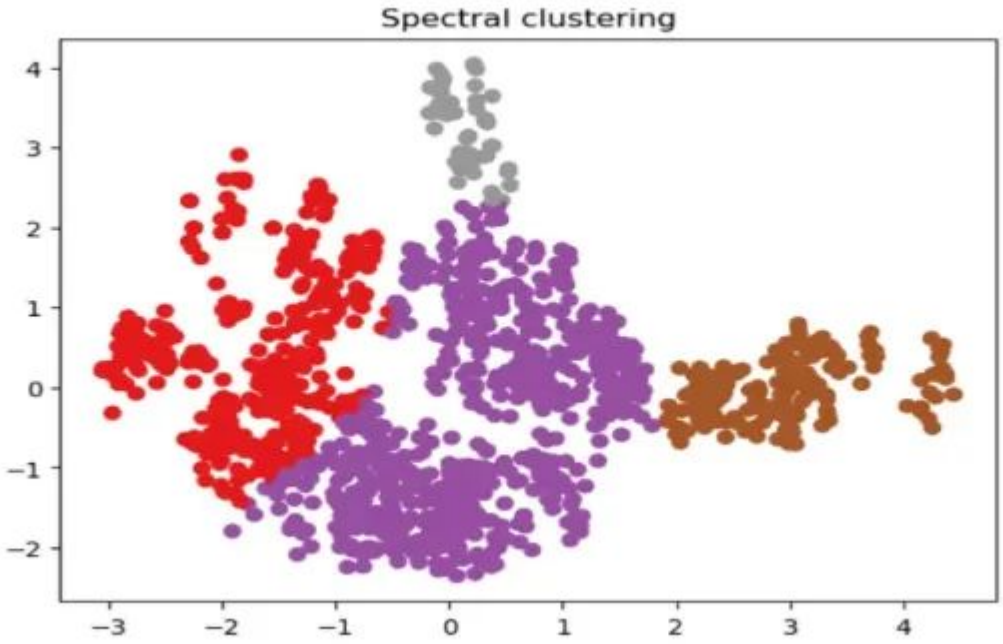
Agrupamiento espectral

La agrupación espectral es una técnica que se utiliza para agrupar entidades similares (en este caso, nodos de un grafo) en clústeres o grupos. En lugar de basarse únicamente en las conexiones directas entre nodos, utiliza la matriz laplaciana, que resume el grafo y representa matemáticamente cómo se conectan los nodos.

Así es como funciona en términos simples:

Agrupamiento espectral

1. La agrupación espectral convierte un gráfico en una matriz y analiza sus valores propios (números especiales que capturan la estructura del gráfico).
2. Estos valores propios ayudan al algoritmo a comprender la “forma” general del gráfico.
3. Basándose en esta información, el algoritmo agrupa nodos similares en clústeres.



Agrupamiento espectral

Es especialmente útil para grafos complejos cuya estructura no es sencilla. Por ejemplo, en un grafo disperso (donde muchos nodos no están conectados directamente), la agrupación espectral permite identificar grupos significativos considerando las relaciones indirectas entre nodos.

En resumen, la agrupación espectral utiliza matemáticas —específicamente la matriz laplaciana y los valores propios— para agrupar nodos similares, incluso si no están directamente conectados. Esto la convierte en una herramienta poderosa para comprender redes complejas.

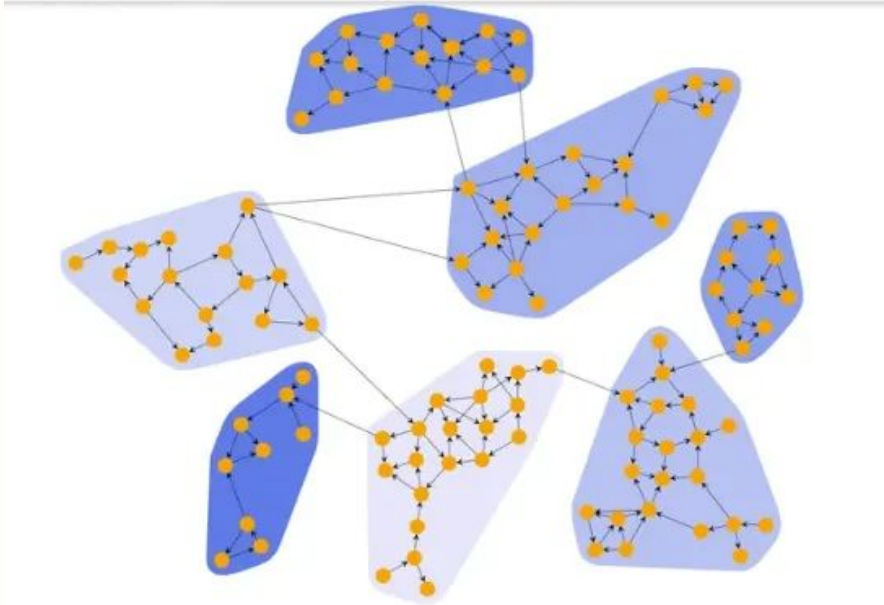
Agrupamiento de intermediación de bordes

Este método permite encontrar comunidades o grupos en una red (como personas en una red social o dispositivos en una red de comunicación). Funciona observando los bordes (conexiones entre nodos) y centrándose en aquellos que son más importantes para conectar las diferentes partes del grafo. Esta importancia se mide mediante la centralidad de intermediación, que indica la frecuencia con la que un borde en particular actúa como atajo o puente entre diferentes grupos de nodos.

Así es como funciona el método paso a paso:

Agrupamiento de intermediación de bordes

1. La centralidad de intermediación identifica los bordes que son “críticos” para conectar diferentes partes del gráfico.
2. Los bordes con alta intermediación se eliminan uno por uno, cortando esencialmente el gráfico en pedazos.
3. Al repetir este proceso, el gráfico se divide gradualmente en grupos o comunidades más pequeños donde los nodos dentro de cada grupo están más conectados entre sí que con los nodos de otros grupos.



Agrupamiento de intermediación de bordes

Este enfoque es excelente para encontrar estructuras comunitarias en redes, donde buscamos grupos de nodos con conexiones internas más estrechas, pero menos estrechas con otros grupos. Es especialmente útil para detectar comunidades estructurales donde los nodos están interconectados de forma compleja.

Agrupamiento basado en redes neuronales gráficas

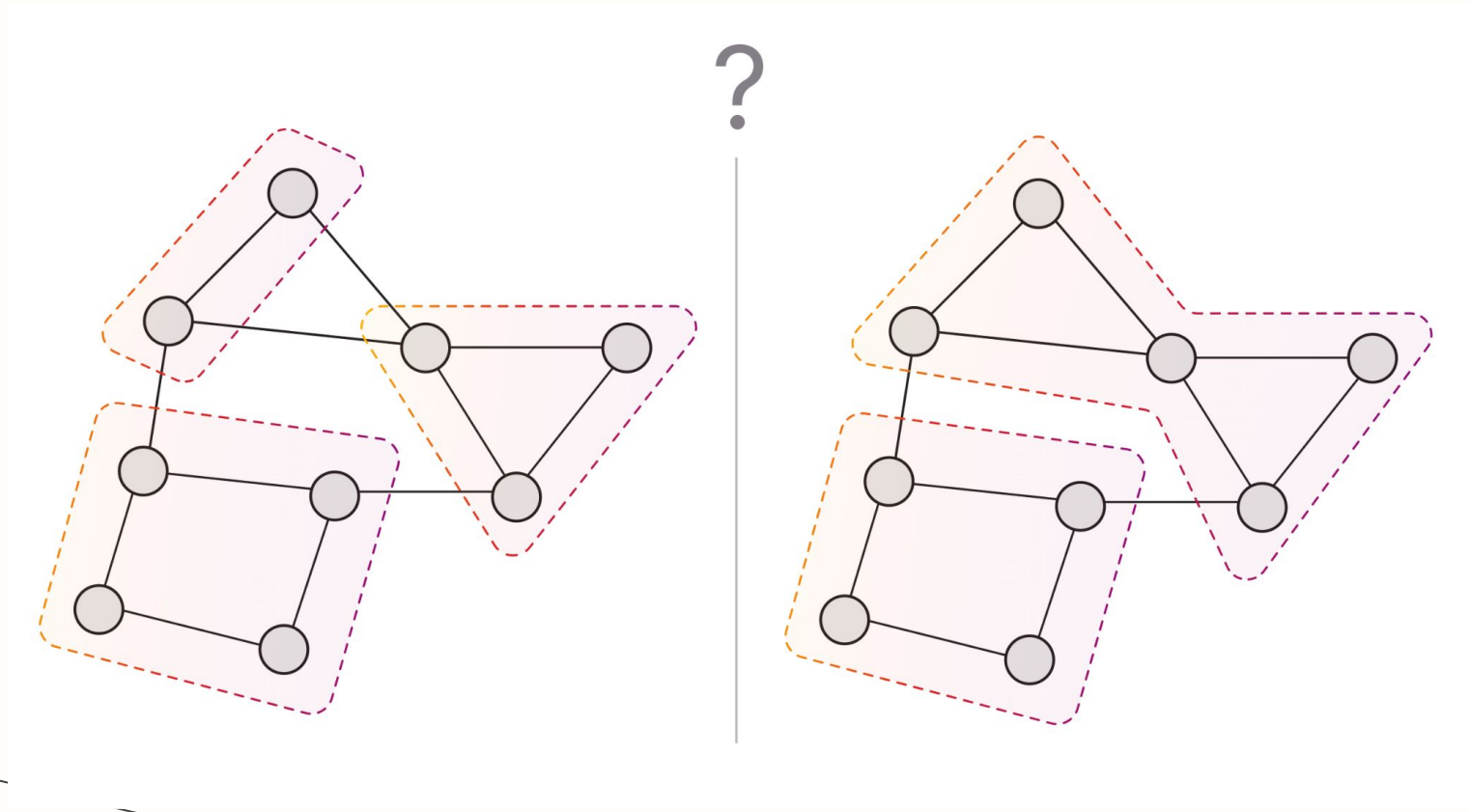
Las redes neuronales de grafos (GNN) se utilizan principalmente para la clasificación de nodos, la predicción de aristas y la clasificación de grafos, pero también pueden adaptarse para la agrupación en clústeres no supervisada. Si bien las GNN no realizan la agrupación en clústeres directamente, sus técnicas de incrustación de nodos proporcionan representaciones que pueden agruparse mediante métodos tradicionales como k-medias o técnicas avanzadas como la agrupación en clústeres contrastiva.

Varios enfoques de agrupamiento de gráficos profundos aprovechan las GNN:

Agrupamiento basado en redes neuronales gráficas

1. **Autocodificadores de gráficos (GAE, VGAE):** utilizan codificadores GNN para aprender representaciones de nodos latentes, que luego pueden agruparse mediante agrupamiento espectral o k-medias.
2. **Agrupamiento de gráficos contrastivos:** técnicas como AGE, MVGRL y HeCo refinan las incrustaciones de nodos mediante aprendizaje contrastivo, lo que hace que los grupos sean más separables.
3. **Agrupación de gráficos para agrupamiento:** métodos como DiffPool y MinCutPool agregan características de nodos para crear estructuras de agrupamiento jerárquicas dentro del gráfico.

Elegir el algoritmo adecuado



Factores específicos del dominio

- **Agrupamiento jerárquico:** adecuado para conjuntos de datos más pequeños o aplicaciones donde la interpretabilidad es fundamental, como conjuntos de datos biológicos.
- **Algoritmos basados en modularidad:** efectivos para la detección de comunidades en redes sociales y bioinformática.
- **Propagación de etiquetas:** Escala bien para gráficos grandes (de millones a miles de millones de nodos), como los gráficos web, donde los nodos son páginas web y los bordes son hipervínculos. Se utiliza para la detección de spam y la clasificación de temas.

Factores específicos del dominio

- **Agrupamiento espectral:** captura estructuras de clústeres complejas en gráficos dispersos o ponderados, como redes de transporte, donde los flujos de tráfico unen estaciones.
- **Agrupamiento de intermediación de bordes:** detecta comunidades bien separadas en redes más pequeñas (hasta decenas de miles de nodos), como redes académicas o corporativas.

Factores Técnicos

- **Eficiencia computacional:** Los métodos espectrales y modulares pueden requerir un alto consumo computacional para grafos muy grandes. La propagación de etiquetas ofrece un mejor rendimiento en estos casos.
- **Requisitos de memoria:** Los métodos de incrustación de nodos requieren una cantidad significativa de memoria para el almacenamiento de vectores, pero brindan resultados más completos.
- **Complejidad de los gráficos:** la agrupación espectral se destaca en el manejo de gráficos dispersos o ponderados, mientras que los métodos basados en modularidad son más adecuados para gráficos no dirigidos.

Factores Técnicos

Agrupamiento jerárquico:

- Tiene dificultades para escalar grandes conjuntos de datos debido a la alta sobrecarga computacional.
- Ofrece alta interpretabilidad, lo que lo hace adecuado para conjuntos de datos pequeños y estructurados.
- Requiere un tiempo de procesamiento significativo para gráficos densamente conectados.

Algoritmos basados en modularidad:

- Computacionalmente intensivo para gráficos más grandes, especialmente con alta densidad de bordes.
- Funciona bien para gráficos no dirigidos con estructuras comunitarias claras.
- Menos adecuado para gráficos ponderados o dirigidos sin modificaciones.

Factores Técnicos

Propagación de etiquetas:

- Se escala de manera eficiente para conjuntos de datos muy grandes.
- Fácil de implementar y computacionalmente liviano.
- Carece de un control detallado sobre el tamaño y los límites del clúster.

Agrupamiento espectral:

- Computacionalmente costoso debido a la descomposición de valores propios.
- Eficaz para gráficos dispersos y ponderados con estructuras complejas.
- Requiere un número predefinido de clústeres, lo que limita la adaptabilidad para tareas exploratorias.

Agrupamiento de intermediación de bordes:

- Ineficiente para gráficos de gran escala debido a la necesidad de volver a calcular la centralidad del borde.
- Se destaca en la detección de comunidades bien separadas y distintas.
- No es ideal para gráficos densos o altamente interconectados.

Consideraciones prácticas

Tamaño del gráfico
Los algoritmos como la propagación de etiquetas son eficientes para gráficos muy grandes.

Interpretabilidad de los resultados.
Los enfoques basados en la modularidad a menudo producen grupos que son fáciles de interpretar.

Características de los datos
Para gráficos ponderados o dirigidos, los algoritmos especializados pueden funcionar mejor.

Consideraciones prácticas

Necesidades de escalabilidad
Para las aplicaciones en tiempo real, la eficiencia computacional se vuelve primordial.

Facilidad de implementación
La propagación de etiquetas es sencilla de implementar e interpretar, lo que la hace práctica para una agrupación rápida.

Interpretabilidad
La agrupación jerárquica proporciona una estructura clara y fácil de explicar, mientras que los enfoques basados en la modularidad también ofrecen una buena interpretabilidad.

Casos de uso especializados
La agrupación espectral es ventajosa para gráficos ponderados o dirigidos, mientras que los GNN son versátiles para aplicaciones que requieren adaptabilidad dinámica.