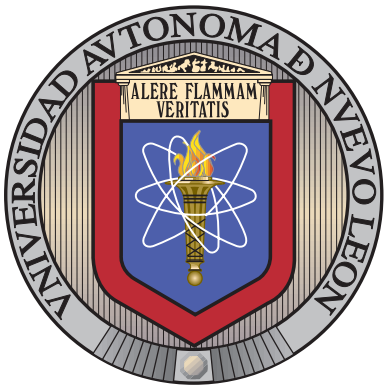
Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

División de Estudios de Posgrado



Agrupamiento Local en Grafos Dirigidos

por

Vanesa Avalos Gaytán

en opción al grado de

Maestría en Ciencias

en Ingeniería de Sistemas

San Nicolás de los Garza, Nuevo León febrero 2009

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

División de Estudios de Posgrado

Agrupamiento Local en Grafos Dirigidos

por

Vanesa Avalos Gaytán

en opción al grado de

Maestría en Ciencias

en Ingeniería de Sistemas

San Nicolás de los Garza, Nuevo León febrero 2009

División de Estudios de Posgrado

Los miembros del Comité de Tesis recomendamos que la Tesis *Agrupamiento Local en Grafos Dirigidos*, realizada por la alumna Vanesa Avalos Gaytán, con número de matrícula 1437964, sea aceptada para su defensa como opción al grado de Maestría en Ciencias en Ingeniería de Sistemas.

El Comité de Tesis

Dra. Satu Elisa Schaeffer
Asesor

Dr. Igor S. Litvinchev
Revisor

Dr. Humberto Madrid de la Vega
Revisor

Vo. Bo.

Dr. Moisés Hinojosa Rivera
División de Estudios de Posgrado

San Nicolás de los Garza, Nuevo León, febrero 2009

Dedicatoria

A mamá y papá que tanto quiero.

Agradecimientos

Gracias a la coordinación del Programa Posgrado en Ingeniería de Sistemas (PISIS) por darme la oportunidad de formar parte de éste. A la Universidad Autónoma de Nuevo León y a la Facultad de Ingeniería Mecánica y Eléctrica por el apoyo que me han dado durante la realización de la maestria. Al Consejo Nacional de Ciencia y Tecnología CONACYT por la beca de manutención otorgada durante mis estudios de maestría.

Mi más sincero agradecimiento a la Dra. Satu Elisa Schaeffer con quien he llevado acabo la realización de éste trabajo, porque siempre tiene tiempo para escucharme, resolver mis dudas y sobre todo por entenderme y tenerme mucha paciencia.

A los miembros del comité de tesis, Dra. Satu Elisa Schaeffer y Dr. Humberto Madrid de la Vega, por sus constantes recomendaciones para el mejoramiento de éste trabajo; gracias al Dr. Igor Litvinchev por formar parte del comité.

A todos los profesores del PISIS por su apoyo durante mi estancia en el PISIS.

Este trabajo ha recibido apoyo de los proyectos de la Universidad Autónoma de Nuevo León PAICyT CA1475-07 y la Secretaría de Educación Pública PROMEP 103,5/2523/07.

Resumen

Vanesa Avalos Gaytán.

Candidato para el grado de Maestra en Ciencias en Ingeniería
con especialidad en Ingeniería de Sistemas.

Universidad Autónoma de Nuevo León.
Facultad de Ingeniería Mecánica y Eléctrica.

Título del estudio:

Agrupamiento Local en Grafos Dirigidos

Número de páginas: 57.

En el presente trabajo nos enfocamos en la realización de una nueva técnica para el agrupamiento local en grafos dirigidos mediante el uso de los tiempos de absorción y caminatas aleatorias de una cadena de Markov.

Nos hemos enfocado en el agrupamiento local ya que actualmente se sabe que el trabajo existente para agrupamiento local es escaso, por otro lado también sabemos que para el agrupamiento global en grafos existen métodos que dan buenos resultados. Sin embargo, estos métodos son muy costosos computacionalmente cuando se tiene un grafo de tamaño masivo y no se tiene interés en conocer todos los grupos existentes en el grafo. La forma que proponemos para hacer agrupamiento local hace uso de resultados existentes.

El principal resultado de este trabajo es que se ha logrado obtener una buena medida para hacer agrupamiento local por medio de los tiempos de absorción y al hacer la comparación con el agrupamiento local obtenido por caminatas aleatorias, los resultados han sido los que se esperaban.

Como trabajo futuro se propone extender los métodos a grafos ponderados para hacer agrupamiento local ya que los resultados existentes para agrupamiento global son aplicables a grafos simples y ponderados.

Dra. Satu Elisa Schaeffer

Asesor

Estructura de la tesis

En el presente trabajo se aborda el problema de *Agrupamiento local en grafos dirigidos*, el problema y los conceptos descritos se presentan de la forma más simple posible para un mejor entendimiento de nuestro trabajo.

En el capítulo 1 se describe lo que es el agrupamiento de datos y se mencionan algunos trabajos desarrollados. También exponemos cúal es el objetivo de esta tesis así como la hipótesis planteada y la metodología utilizada para el desarrollo del presente trabajo. En el capítulo 2 introducimos conceptos básicos sobre teoría de grafos usados durante el trabajo que desarrollamos.

En el capítulo 3 nos enfocamos en describir qué es el agrupamiento de grafos, cúal es su objetivo y mencionamos ejemplos. Describimos el trabajo existente que se ha desarrollado años atrás sobre métodos espectrales para agrupamiento, en particular para agrupamiento global y mostramos con un ejemplo en que consiste. Introducimos en qué consiste el agrupamiento local y cúal es su objetivo.

En el capítulo 4 se introducen propiedades relacionadas con *cadenas de Markov* ya que la solución que proponemos para hacer agrupamiento local en grafos dirigidos se basa en el uso de los tiempos de absorción y caminatas aleatorias en cadenas de Markov. De la misma manera, introducimos notación matemática para plantear nuestro problema sin ningun inconveniente y vemos cómo el problema de partición espectral puede ser modelado como un problema de programación entera.

En el capítulo 5 mostramos la solución propuesta, finalmente en el capítulo 6 explicamos la solución propuesta por medio de caminatas aleatorias y damos los

resultados obtenidos, en el capítulo 7 damos algunas conclusiones.

Índice general

Agradecimientos v

Resumen vi

Estructura de la tesis viii

1. Introducción 1

1.1. Agrupamiento de datos 1

1.2. Objetivo 4

1.3. Hipótesis 5

1.4. Metodología 5

2. Teoría de grafos 8

2.1. Definiciones 8

2.2. Matrices: incidencia, adyacencia y Laplace 11

3. Trabajo existente 15

3.1. Agrupamiento de grafos 15

3.2. Métodos espectrales para agrupamiento 17

3.3. Agrupamiento global de grafos 17

3.4. Agrupamiento local de grafos 19

4. Cadenas de Markov **21**

4.1. Caminatas aleatorias 22

4.2. Partición espectral como un problema de relajación de un programa entero 27

5. Vector de Fiedler local y tiempos de absorción **30**

5.1. Aproximación local del vector de Fiedler 33

6. Agrupamiento por caminatas aleatorias **35**

6.1. Experimentos de caminatas aleatorias 39

7. Conclusiones **45**

Capítulo 1

Introducción

1.1 Agrupamiento de datos

El *agrupamiento* es la clasificación de objetos en diferentes grupos, ó la *partición* de un *conjunto de datos* en *subconjuntos* (inglés: *clusters*) de modo que cada subconjunto comparte alguna propiedad en común, por ejemplo: color, textura o tamaño, ver Figura 1.1.

Generalmente, los objetos son agrupados por la proximidad de acuerdo a alguna medida de distancia definida, por ejemplo si se tiene un conjunto de puntos en el espacio, usualmente la distancia utilizada es la distancia Euclideana entre dos puntos de este conjunto.

Existen medidas para medir que tan bueno es un agrupamiento, por ejemplo la

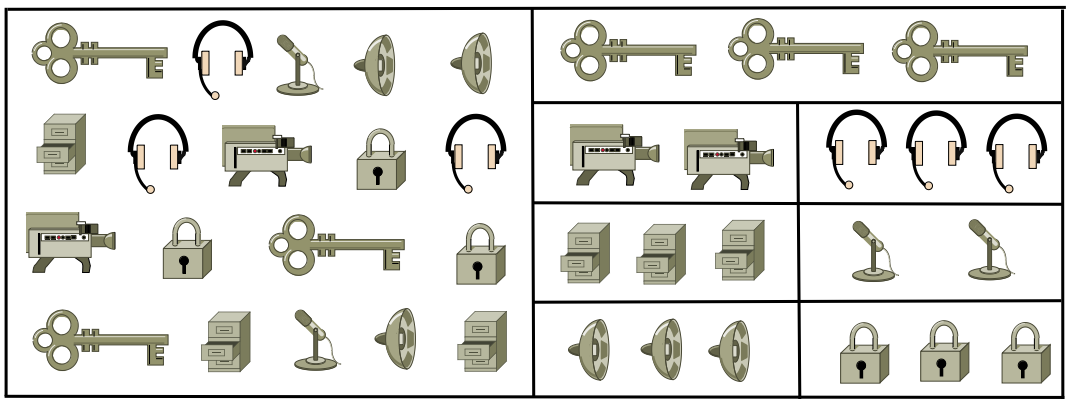


Figura 1.1: Diferentes objetos clasificados por propiedades en común.

medida F [44] la cual necesita de dos agrupamientos, uno obtenido mediante algún método y el otro realizado manualmente. Ésta da información sobre que tan relacionados están los agrupamientos; cuanto más cercana sea a uno, más relacionados están los agrupamientos. La medida de Dunn [17] es la razón entre la distancia mínima dentro de un grupo y la distancia máxima entre los grupos, esta medida da valores altos para agrupamientos con grupos compactos y bien separados. Sin embargo, se sabe que está medida es muy sensible [6]. Otra medida que se usa es el índice de Davies-Bouldin [14], éste es la proporción entre la suma de la medida de dispersión dentro de cada grupo y la separación entre ellos. La medida de dispersión se basa en la suma de la distancia entre el centro del grupo y el resto de sus elementos. Valores pequeños de este índice corresponden a grupos que son compactos y cuyos centros están alejados uno del otro. Existen otras medidas que se basan en medir la densidad de la relación entre los elementos de un conjunto de datos [44].

Agrupamiento de datos es un campo de investigación con numerosas aplicaciones [25]. Es una técnica común en *análisis estadístico de datos*, el cual es utilizado en muchos campos, incluyendo *minería de datos, reconocimiento de patrones, análisis de imágenes* y *bioinformática*. El objetivo es descubrir grupos de elementos altamente relacionados en un gran conjunto de datos; muchos de éstos permiten una representación natural en forma de *grafos*. En la actualidad existen varios traba-

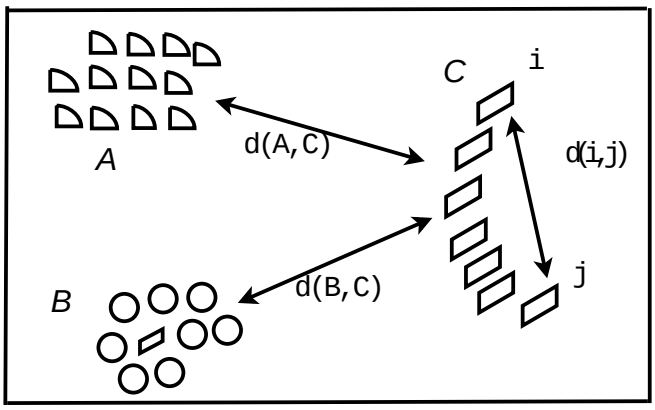


Figura 1.2: Un agrupamiento de tres grupos, $d(A, C)$ y $d(B, C)$ denotan la distancia entre los grupos y $d(i, j)$ es la distancia entre objetos de un grupo.

jos acerca de agrupamiento de grafos *no dirigidos* [39], muchos de ellos utilizando *técnicas espectrales* de grafos [10], con aplicaciones sobre segmentación de imágenes usando agrupamiento de grafos [40]. En los últimos años se han desarrollado algunos métodos de agrupamiento para grafos dirigidos. Por ejemplo, Meila y Pentney [30] abordan el agrupamiento para grafos dirigidos utilizando una variante del algoritmo de Shi y Malik [40].

Los métodos mencionados se basan en técnicas espectrales de grafos y agrupan todo el grafo al mismo tiempo, procesando información *global* de los datos, lo cual es muy costoso computacionalmente cuando no se tiene interés en conocer todos los grupos existentes. El trabajo que nosotros deseamos hacer es desarrollar métodos de agrupamiento *local* con los cuales se puedan obtener agrupamientos buenos para grafos dirigidos de tamaño masivo.

Entre los trabajos más recientes dedicados al agrupamiento local de grafos podemos mencionar los de Chung [8], por ejemplo trabajos en los que estudia los valores propios de matrices asociadas a grafos dirigidos, analiza la conexión con el cociente de Rayleigh y utiliza la constante de *Cheeger* para establecer la desigualdad de Cheeger para grafos dirigidos. Las relaciones que pueden existir entre lo anteriormente mencionado pueden servir para enfrentar diferentes problemas que surgen a menudo en el estudio de cadenas de Markov [8].

Entre otros de los trabajos de Chung que podemos destacar son trabajos relacionados con agrupamiento local para grafos dirigidos usando el PageRank [3] y caminatas aleatorias [11], en los que se hace uso de la constante de Cheeger, propiedades de caminatas aleatorias, la conductancia de grafos y métodos espectrales.

Un *conjunto de datos* es una colección de datos, usualmente presentados dentro de una tabla. Cada columna representa una variable particular. Cada fila corresponde a un miembro dado del conjunto de datos en cuestión. En esta tabla se enumeran los valores para cada una de las variables, por ejemplo altura y peso de un objeto. En matemáticas, distancia se caracteríza como una función que mide la disimilitud

entre elementos de un conjunto, como se muestra en la Figura 1.2.

La *minería de datos* [5] es el principio de los procesos de selección a través de grandes cantidades de datos y obtener la información pertinente. Usualmente es utilizado por organizaciones de inteligencia de negocios y analistas financieros, pero es cada vez más utilizado en las ciencias para extraer información de los enormes conjuntos de datos generados por los métodos experimentales modernos y de observación. La minería de datos se ha descrito como "extracción no trivial de la implícita, previamente desconocida y potencialmente útil información de los datos" y "la ciencia de extraer información útil de grandes conjuntos de datos o bases de datos" [5]. El análisis de datos es el proceso de examinar datos con el objetivo de extraer solamente los que aportan mayor información con el fin de llegar a alguna conclusión razonablemente buena sin tener que utilizar todos los datos.

El análisis de datos está muy relacionado con la minería de datos. Ésta tiende a centrarse en los conjuntos de datos más grandes pero con menos énfasis en hacer inferencias; a menudo utiliza los datos originales para un propósito diferente. Frecuentemente, el análisis de datos se divide en *análisis exploratorio de datos* (AED) y *análisis de datos de confirmación* (ADC) [5]. El objetivo del AED es descubrir nuevas características en los datos mientras que el objetivo del ADC es la confirmación o falsificación de las hipótesis existentes.

1.2 Objetivo

Nuestro objetivo es desarrollar investigación básica para problemas de *agrupamiento de grafos* con el fin de lograr un mayor entendimiento sobre las técnicas de agrupamiento y entender a profundidad la estructura matemática del problema que planteamos para desarrollar una técnica de solución nueva, la cual sea eficiente y justificada matemáticamente. Así mismo, también deseamos que la técnica desarrollada explote tal estructura favorablemente y que posteriormente permita otras variaciones más.

La técnica desarrollada es implementada por medio de un algoritmo para *agrupamiento local en grafos dirigidos*, al mismo tiempo que nos basamos en los métodos ya existentes para grafos no dirigidos.

1.3 Hipótesis

La hipótesis que nos planteamos es que partiendo de la información existente sobre agrupamiento global de grafos no dirigidos podemos desarrollar e implementar métodos para el agrupamiento local de grafos dirigidos eficientes y justificados matemáticamente.

En el mundo real existen diversos problemas los cuales pueden ser resueltos mediante el análisis de grafos. En la actualidad es de gran interés estudiar problemas como el que trabajaremos ya que los resultados que se desean obtener pueden ser aplicados en diversas áreas como por ejemplo: minería de datos, bioinformática, medicina y comunicación. Existen resultados los cuales son aplicables a problemas modelados mediante grafos no dirigidos. Nosotros deseamos desarrollar métodos para grafos dirigidos a partir de los métodos existentes.

1.4 Metodología

La metodología aplicada es el desarrollo e implementación de algoritmos eficientes de agrupamiento local de grafos dirigidos que aún no existen. Dado el enfoque matemático en el diseño de las medidas de calidad del agrupamiento, el desarrollo y estudio de tales algoritmos pertenece a un área entre las matemáticas y ciencias computacionales. Uno de los objetivos de este trabajo es desarrollar algoritmos conceptuales para el agrupamiento local en grafos de tamaño masivo. No es nuestro objetivo la demostración de teoremas mátematicos ni de los algoritmos, sino hacer uso de las matemáticas y técnicas existentes para llegar a un resultado con el cual se puedan obtener agrupamientos locales.

Las aplicaciones reales provienen de varias ramas de la ciencia. Para llevar acabo exitosamente el desarrollo de ésta investigación, se satisfacen los siguientes puntos:

1. Adquirir un buen entendimiento matemático de la estructura de subgrafos, basado en el álgebra de grafos y propiedades espectrales tales como los valores y vectores propios de las matrices asociadas.

2. Detectar propiedades de calidad para determinar hasta que grado un subgrafo dado forma un buen grupo para un agrupamiento.

3. Aprovechar de estas propiedades para determinar el mejor grupo para un vértice de interes dado.

4. Realizar experimentación aplicada a datos artificiales generados y datos de problemas existentes relacionados con agrupamiento para probar los métodos.

En cada uno de los puntos anteriores prestamos cuidado especial para que la adaptación de los métodos también pueda ser extendida a grafos ponderados. Así mismo debemos explotar totalmente la estructura no uniforme de instancias típicas, en lugar de desarrollar métodos para "casos promedios aleatorios". De este modo podemos garantizar nuevos resultados y métodos útiles para el uso práctico en diferentes áreas de aplicación.

En el caso de grafos ponderados y dirigidos, el problema de agrupamiento local no tiene aún soluciones satisfactorias fundamentadas matemáticamente ni éxito en la práctica. En las aplicaciones por lo general los grafos son ponderados y en muchos de los casos dirigidos. Ambos aspectos dan al problema de agrupamiento local una estructura matemática diferente a los problemas relacionados con grafos no dirigidos, por lo cual es importante el estudio y solución eficiente del desarrollo de métodos eficientes para grafos dirigidos y ponderados. Aunque por lo general serán de tiempo exponencial, ya que la naturaleza del problema de agrupamiento es NP-dura [39].

Así, la *contribución* de éste trabajo es el desarrollo e implementación de métodos eficientes que sean justificados matemáticamente.

Capítulo 2

Teoría de grafos

En este capítulo describimos las definiciones necesarias para familiarizarnos con la teoría de grafos [7, 15], ya que ésta es la base fundamental de nuestro trabajo.

2.1 Definiciones

Grafo: es un par de conjuntos (V, E), donde V es un conjunto finito de puntos v_1, v_2, v_3, ..., v_n llamados vértices o nodos y E es un conjunto finito de aristas e_{ij}, cada uno de los cuales une pares ordenados de vértices. A las aristas se les puede asignar un valor o peso w_{ij}, Figura 2.1.

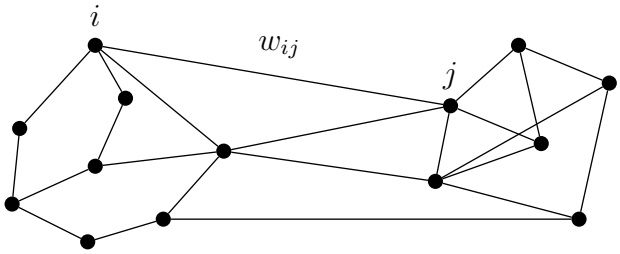


Figura 2.1: Grafo simple.

Grado de un vértice: es la suma (de los pesos) de los aristas incidentes a un vértice. Se denota por $d_i = \sum_j w_{ij}$, Figura 2.2. Con frecuencia se considera:

$$w_{ij} = \begin{cases} 1 & \text{si la arista } e_{ij} \text{ existe} \\ 0 & \text{en otro caso.} \end{cases}$$

En este caso d_i es el número de aristas incidentes en el vértice i.

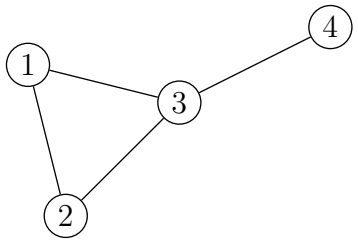


Figura 2.2: El grado para cada uno de los vértices del grafo son: $d_1 = 2$, $d_2 = 2$, $d_3 = 3$, $d_4 = 1$.

Grafo completo: Si cada par de vértices está conectado por un arista, se dice que el grafo es completo, Figura 2.3. Si G tiene n vértices, el número de aristas es $\frac{n(n-1)}{2}$.

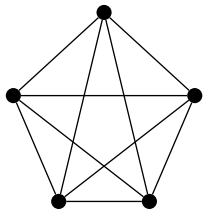


Figura 2.3: En un grafo completo, cada vértice tiene grado $n - 1$.

Camino: Es una sucesión finita en la que aparecen alternadamente vértices y aristas de un grafo dado G.

Grafo conexo: Si existe un camino entre cualesquiera dos vértices se dice que el grafo G es conexo, Figura 2.4.

Constante de Cheeger: La constante de *Cheeger* de un grafo es una medida numérica que es estrictamente positiva si y sólo si G es un grafo conexo. Intuitivamente, si la constante es muy pequeña pero positiva, entonces se dice que existe un cuello de botella en el sentido que hay dos conjuntos de vértices tales

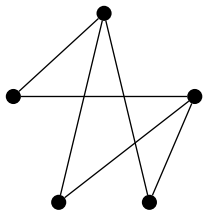


Figura 2.4: En este ejemplo existe un camino entre cada par de aristas

que entre ellos existen pocas aristas pero dentro de tales subconjuntos existen muchas aristas. La constante es grande si existen muchos aristas entre los subconjuntos en que se ha dividido el conjunto de vértices [8, 9].

Grafo dirigido: Las aristas tienen dirección del vértice i (punto inicial) al vértice j (punto final). Así, el grado del vértice i es el número de aristas que tienen punto inicial en el mismo. Los vértices que no tienen aristas con punto inicial en él tienen grado cero como se muestra en la Figura 2.5.

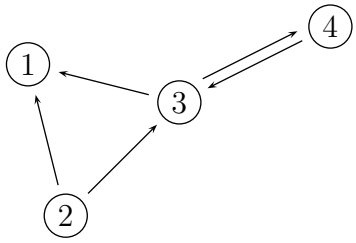


Figura 2.5: Para un grafo dirigido como el que se muestra el grado de los vértices es: $d_1 = 0$, $d_2 = 2$, $d_3 = 2$, $d_4 = 1$.

Grafo no dirigido: Las aristas no tienen dirección; el grado de un vértice es el número de aristas incidentes en sí mismo.

Subgrafo: Sea $A(V_1, E_1)$ un grafo, A es una subgrafo de G si $V_1 \subset V$ y $E_1 \subset E$, Figura 2.6.

Diámetro: La distancia entre dos vértices es el menor número de aristas que se requieren para ir de un vértice i a un vértice j. El diámetro en un grafo es la mayor distancia entre dos vértices.

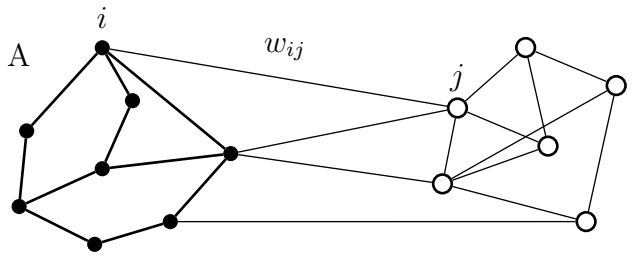


Figura 2.6: Los vértices y aristas en negrita forman el subgrafo A.

Volumen de un subgrafo: Es la suma de los grados de los vértices que pertenecen al subgrafo A. Se denota por $V(A) = \sum_{i \in A} d_i$

Bipartición de un grafo: Si un grafo $G(V, E)$ puede ser dividido en dos subgrafos A y B tales que $A \cup B = V$ y $A \cap B = \emptyset$, eliminando los aristas que unen A y B, entonces se dice que (A, B) es una partición de V.

2.2 Matrices: incidencia, adyacencia y Laplace

Existen diferentes formas de almacenar grafos en una computadora. La estructura de datos usada depende de las características del grafo y el algoritmo usado para manipularlo. Teóricamente se puede distinguir la estructura de listas y las de matrices, pero usualmente, lo mejor es una combinación de ambas. Las listas son preferidas en grafos dispersos porque tienen un eficiente uso de la memoria. Por otro lado, las matrices proveen acceso rápido, pero pueden consumir grandes cantidades de memoria.

Matriz de incidencia: Un grafo no dirigido con n vértices y m aristas tiene asociado una matriz de incidencia I_G. Denotemos por v_i y v_j los vértices incidentas al arista e_{ij}. En I_G en la posición correspondiente a la columna e_{ij} y la fila v_i hay un 1 y en la posición correspondiente a la columna e_{ij} y la fila v_j hay un -1; el resto de los elementos son cero. Por ejemplo en la Figura 2.7 mostramos un grafo conexo de seis vértices y su matriz de incidencia es la ecuación 2.1.

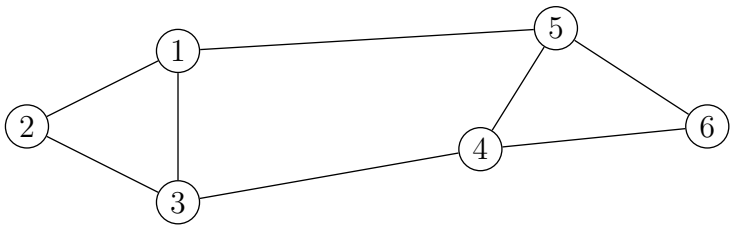


Figura 2.7: Grafo simple conexo de seis nodos y su matriz de incidencia.

$$I_G = \begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{array} \begin{array}{c} \begin{array}{cccccccc} e_{12} & e_{13} & e_{15} & e_{23} & e_{34} & e_{45} & e_{46} & e_{56} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 \end{pmatrix} \end{array}. \tag{2.1}$$

Una o más columnas de la matriz de incidencia pueden ser multiplicadas por -1, así, cada matriz obtenida sigue siendo una matriz de incidencia para el mismo grafo. Por ejemplo:

$$I_G = \begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{array} \begin{array}{c} \begin{array}{cccccccc} e_{12} & e_{13} & e_{15} & e_{23} & e_{34} & e_{45} & e_{46} & e_{56} \end{array} \\ \begin{pmatrix} -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array} \tag{2.2}$$

sigue siendo una matriz de incidencia para el grafo de la Figura 2.7.

Matriz de adyacencia: Si G es un grafo con n vértices la matriz $\mathbf{A}$ $n \times n$, cuyo elemento a_{ij} es 1 si existe la arista e_{ij} y cero en otro caso, es llamada matriz

adyacente de G. La matriz $\mathbf{A}$ para grafo de la Figura 2.7 es:

$$\mathbf{A} = \begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{array} \begin{array}{c} \begin{array}{cccccc} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \end{array}. \tag{2.3}$$

Obsérvese que por definición $\mathbf{A}$ es una matriz simétrica.

Matriz diagonal de grados: Es una matriz diagonal denotada por $\mathbf{D}$, donde el elemento (i, i) es igual al grado del vértice i. En términos de $\mathbf{A}$, $d_i = \sum_j a_{ij}$. Para el grafo de la Figura 2.7 la matriz $\mathbf{D}$ es

$$\mathbf{D} = \begin{pmatrix} 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix}. \tag{2.4}$$

Matriz Laplaciana: Sea G un grafo con n vértices y la matriz adyacente asociada. La matriz Laplaciana de G, denotada por L, es una matriz $n \times n$ con entradas

$$\mathbf{L}_{ij} = \begin{cases} d_{ij} & \text{si } i = j, \\ -a_{ij} & \text{si } i \neq j \end{cases} \tag{2.5}$$

que es lo mismo que $\mathbf{L} = \mathbf{D} - \mathbf{A}$. De esta forma, la matriz $\mathbf{L}$ para el grafo de

la Figura 2.7 es:

$$\mathbf{L} = \begin{pmatrix} 3 & -1 & -1 & 0 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & 0 & 0 & -1 & 3 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix}. \tag{2.6}$$

Matriz Laplaciana Normalizada: Es definida como la matriz $\mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$ y la denotaremos por $\mathcal{L}$. Al normalizar la matriz Laplaciana todos sus valores propios están en el intervalo $[0, 2]$ y el más pequeño es cero [18, 19].

Capítulo 3

Trabajo existente

3.1 Agrupamiento de grafos

Una de las propiedades de interés en el campo de redes naturales es la presencia de grupos (*clusters o comunidades*) [32], esto es, la existencia de subgrafos inducidos densos que tienen relativamente pocas conexiones hacia afuera comparadas con la densidad interna [28].

El objetivo del *agrupamiento de grafos* es agrupar los vértices del grafo en subgrupos tomando en cuenta la estructura de las aristas de forma tal que deben existir *muchas* aristas dentro de los grupos y *relativamente pocas* aristas entre los grupos. En la Figura 3.1 se muestra un grafo formado por conjuntos de vértices que están internamente conectados, los cuales forman pequeños grupos o "cuevas", en los cuales, el último vértice de cada grupo se conecta con el primer vértice del grupo vecino. Los grafos con esta estructura son conocidos como grafos de *hombre de cuevas* [46].

Otro ejemplo clásico es un ejemplo de una pequeña red social del mundo real [48] que frecuentemente es mencionada en la literatura de agrupamiento de grafos [33, 35, 47]. Está es una red social de un club de karate que se dividió en dos grupos (ver Figura 3.2), siendo este un caso ideal para algoritmos de dos–clasificación que agrupa un conjunto de datos u objetos en dos conjuntos. Existen trabajos [39] en los que podemos encontrar información sobre los algoritmos de agrupamiento de grafos.

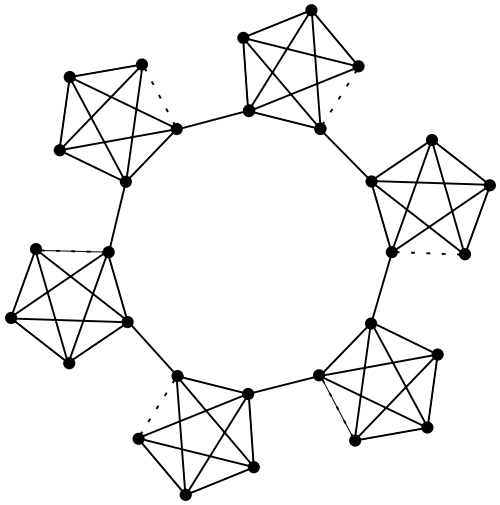


Figura 3.1: Un grafo con estructura de cuevas [46] compuesto de seis grupos los cuales están formados por cinco vértices cada uno, y que están conectados en un grafo circular el cual se forma al "remover" una arista de cada grupo para usarla como la arista que conecta al grupo con su grupo vecino. La arista que se remueve se muestra con una línea punteada.

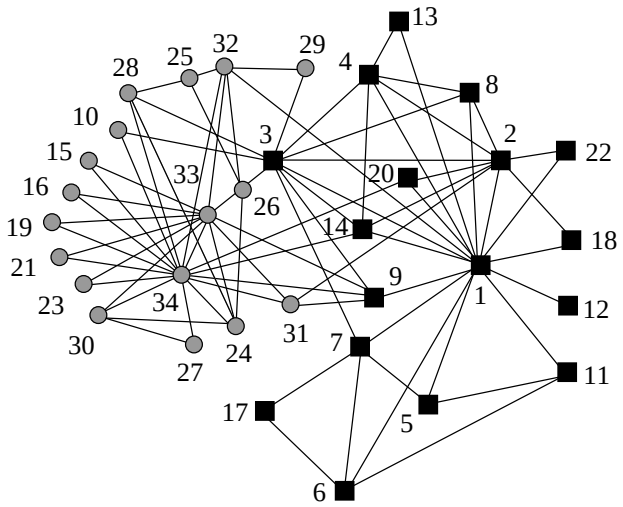


Figura 3.2: Red social del club de karate [48]. Los dos grupos en que se dividió el club son indicados por la forma en que los vértices están dibujados: los cuadros representan un grupo y los círculos otro.

3.2 Métodos espectrales para agrupamiento

El agrupamiento espectral de puntos en el espacio, con frecuencia modelados como grafos ponderados, es un tema ampliamente estudiado [23, 26]. En el contexto de grafos, la técnica usualmente aplicada consiste en que el vector propio derecho asociado al valor propio más pequeño diferente de cero $\mu^{\mathbf{L}}_{(1)}$ de $\mathbf{L}$ es usado para producir una bipartición del grafo tal que los vértices que corresponden a valores negativos en el vector propio forman la bipartición S y los vértices correspondientes a valores positivos están en $S\backslash V$. Este vector propio es llamado *vector de Fiedler*; la técnica fue propuesta por primera vez en 1975 [18, 19]. Para la matriz Laplaciana normalizada el vector correspondiente es llamado vector de Fiedler *normalizado*. Los trabajos sobre el vector de Fiedler basado en agrupamiento espectral son numerosos desde hace varias décadas [1, 24, 37, 43].

3.3 Agrupamiento global de grafos

El *objetivo del agrupamiento global* es determinar en un grafo dado G los grupos existentes. Para fines prácticos es muy caro computacionalmente cuando se tiene interés en determinar solamente el grupo al cual pertenece un vértice.

Por ahora nos enfocaremos en mostrar de la forma más simple en que consiste el agrupamiento global. Se ha mostrado a detalle [20] de manera muy sencilla que dado un grafo G, con el vector propio asociado al segundo valor propio λ_2 más pequeño de la matriz Laplaciana normalizada se puede obtener una buena bipartición:

$$\lambda_2 = \frac{y_2^t \mathbf{D}^{-\frac{1}{2}} \mathbf{L} \mathbf{D}^{-\frac{1}{2}} y_2}{y_2^t y_2}, \tag{3.1}$$

donde y_2 es el vector propio correspondiente de λ_2.

Los signos de los elementos de y_2 determinan los grupos G_1 y G_2, donde $i \in G_1$ si $y_i > 0$ e $i \in G_2$ si $y_i < 0$. El valor propio λ_2 es conocido como el *valor propio de Fiedler* y y_2 como el *vector propio de Fiedler normalizado*. En ocaciones los grupos

determinados por G_1 y G_2 no son los mejores, pues existen casos en los que los signos de y_2 no son el factor determinante para decidir qué vértices forman G_1 y cuáles G_2.

Para el grafo que mostramos en la Figura 3.1, la bipartición basada en $\mathbf{L}$ agrupa tres de las cuevas completas en S de tal forma que asigna valores positivos para cada una de las tres cuevas en $S\backslash V$. Sin embargo, usando el vector propio de Fiedler normalizado, los signos de éste cambian y esto no nos da una división más intuitiva del agrupamiento de las cuevas, es en casos como este en los que debemos tomar en cuenta la estructura del agrupamiento obtenido. Por ejemplo en la Figura 3.3 el agrupamiento con el vector de Fiedler normalizado no es el mejor si consideramos solamente los signos de y_2, sin embargo, observando la estructura que presenta el agrupamiento podemos ver que los mejores grupos son determinados por los conjuntos de vértices marcados en circulos negros y los vértices marcados en circulos blancos. Es por ello que para decidir cual es la mejor bipartición también debemos tomar en cuenta la esctructura que ésta presenta. Los dos vectores se visualizan en la Figura 3.3.

Si sólo existen dos grupos naturales en el grafo, la bipartición con el vector de

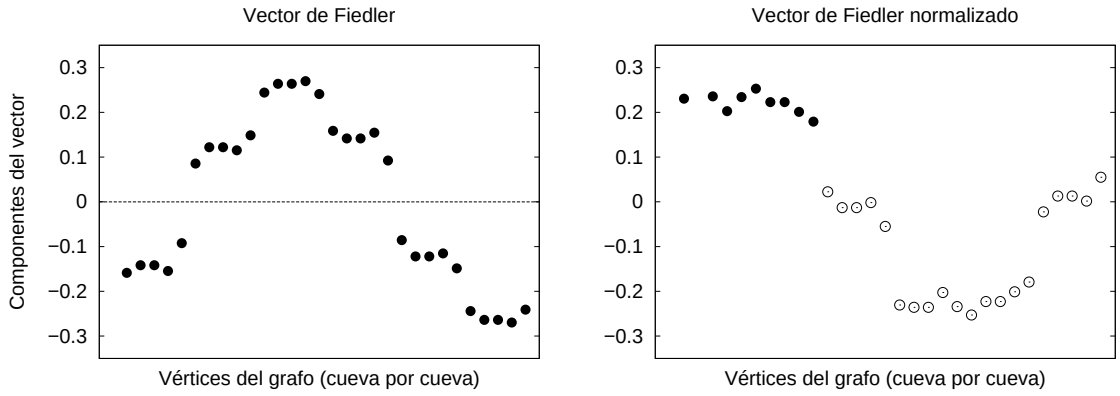


Figura 3.3: Componentes del vector de Fiedler (a la izquierda) y el vector de Fiedler normalizado (a la derecha) para el grafo de cuevas de la Figura 3.1. A simple vista la estructura de los seis grupos es evidente en el vector de Fiedler; en cambio, el vector de Fiedler normalizado agrupa los vértices en cuatro grupos, dos de ellos formados por dos cuevas.

Fiedler está bien. Un ejemplo es la red del club de karate de Zachary de la Figura 3.2: el vector de Fiedler correspondiente es mostrado en la Figura 3.4. Así mismo, el cálculo recursivo de la bipartición en los grupos inducidos por S y $V\backslash S$ ayudará al grafo de entrada a ser agrupado en más de dos grupos, pero para ésto se necesita imponer un *criterio de parada* para determinar cuando terminará la bipartición de los grupos resultantes.

3.4 Agrupamiento local de grafos

Como hemos mencionado antes, el agrupamiento local de grafos se enfoca en determinar el grupo de un vértice de interés también llamado *vértice semilla*. Por ejemplo en la Figura 3.5 podemos observar que el vértice semilla está marcado en amarillo en el grafo. En el agrupamiento local [35] el *objetivo* es encontrar el grupo al cual pertenece un vértice semilla $s \in V$ de interés. Esencialmente, la tarea es encontrar una partición de un grafo G en dos conjuntos de vértices S y $V\backslash S$ tal que $s \in S$ y que S sea un buen grupo en algún sentido predefinido. Comunmente se incluye el criterio de calidad de capacidad de corte y medidas relacionadas tales como

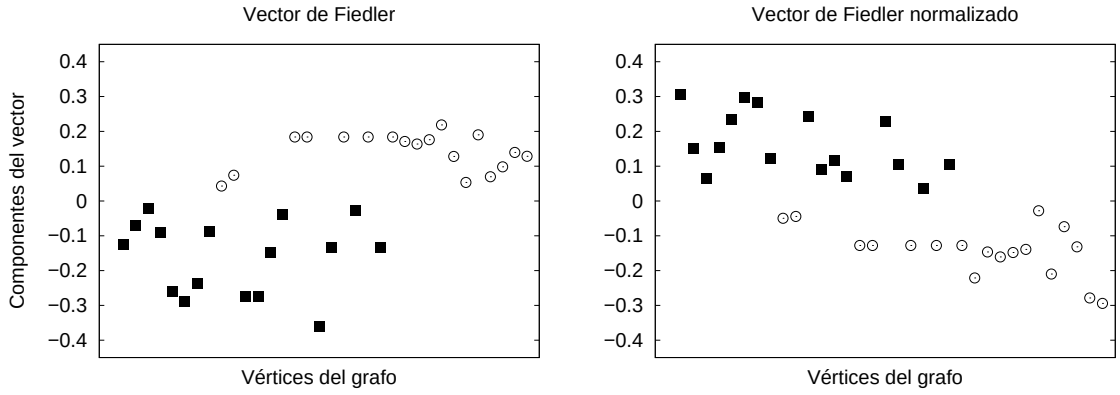


Figura 3.4: Los vértices estan graficados en orden correspondiente a los vértices en la Figura 3.2. Las componentes del vector de Fiedler (a la izquierda) y el vector de Fiedler normalizado (a la derecha) para la red del club de karate de la Figura 3.2. Los vértices pueden ser clasificados en dos grupos: aquellos cuyo valor es positivo en el vector de Fiedler y aquellos cuyo valor es negativo.

conductancia [41] que mide que tan conexo es un grafo en el sentido que controla que tan rápido converge una caminata aleatoria a una distribución uniforme, o medidas basadas en la densidad [38]. También se han propuesto métodos motivados por redes eléctricas tanto para agrupamiento global como para agrupamiento local [34, 35, 47].

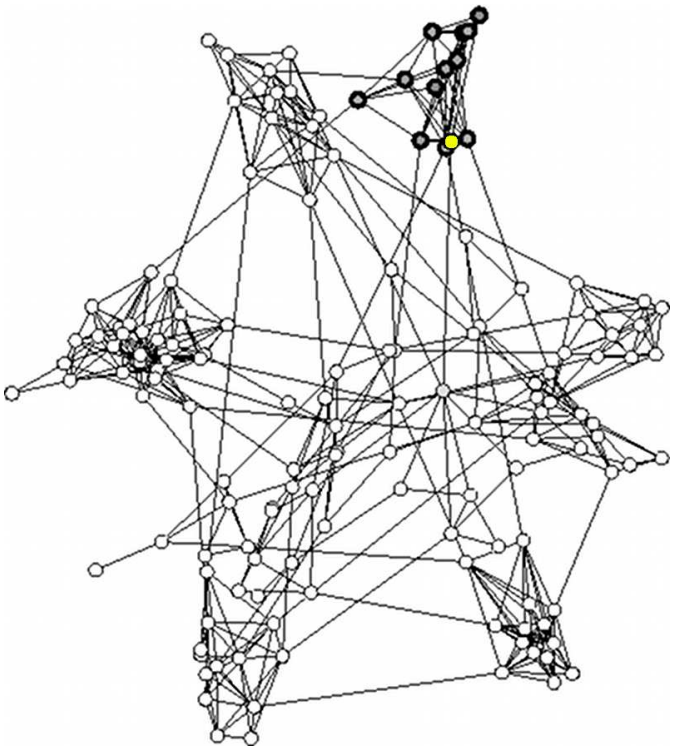


Figura 3.5: El objetivo del agrupamiento local es determinar el grupo al que pertenece un vértice de interés [35], por ejemplo en este grafo en amarillo denotamos el vértice de interés y los vértices con el borde más negro son los que pertenecen a su grupo.

Capítulo 4

Cadenas de Markov

Una cadena de Markov es un *proceso estocástico* en el cual la probabilidad de que ocurra un evento futuro depende sólo del estado actual. Las cadenas de este tipo tienen *memoria*, es decir, "recuerdan" el último evento y esto condiciona las posibilidades de estados futuros, está característica es llamada *Propiedad de Markov* [29]:

$$P\{X_k = i_k | X_0 = i_0, X_1 = i_1, \ldots, X_{k-1} = i_{k-1}\} = P\{X_k = i_k | X_{k-1} = i_{k-1}\}. \quad (4.1)$$

La probabilidad de transición del estado actual i a un estado futuro j es $p(i, j)$, donde

$$p(i, j) = P\{X_k = i_k | X_{k-1} = i_{k-1}\}. \quad (4.2)$$

La *matriz de probabilidad de transición* $\mathbf{P}$ para una cadena de Markov es una matriz $n \times n$, para la cual las entradas $P_{ij} = p(i, j)$. $\mathbf{P}$ es una matriz estocástica, es decir, $0 \leq P_{ij} \leq 1$ para $1 \leq i, j \leq n$, y $\sum_{j=1}^{n} P_{ij} = 1$ para $1 \leq i \leq n$.

En general, cada cadena de Markov, independientemente de cómo sean definidas las probabilidades de transición, puede ser representada por un *grafo dirigido ponderado* donde cada estado en la cadena corresponde a un vértice y cada transición que tiene probabilidad diferente de cero corresponde a una arista y la probabilidad de transición corresponde al peso de la arista. Para un grafo no ponderado, cuando nos movemos de un vértice a otro eligiendo un vértice vecino uniformemente al azar, la matriz de transición que resulta es la matriz de adyacencia normalizada ($\mathbf{D}^{-1}\mathbf{A}$) del grafo G. Esto significa que la probabilidad de moverse del vértice i al vértice j

es simplemente $1/d(i)$.

Un estado j es *accesible* desde cualquier estado i, $(i \to j)$, si existe un entero $k \geq 0$ tal que $P(X_k = j | X_0 = i) > 0$. Cuando $k = 0$, cada estado es definido como accesible desde sí mismo. Se dice que el estado i está *comunicado* con el estado j, $(i \leftrightarrow j)$, si $(i \to j)$ y $(j \to i)$. Un conjunto de estados C es una *clase comunicada* si cada par de estados en C se comunica con cualquier otro par de estados en C. Una clase comunicada es *cerrada* si la probabilidad de salir de la clase es cero, es decir, si $i \in C$ pero $j \notin C$, j no es accesible desde i.

Una cadena de Markov es *irreducible* si su espacio de estados es una clase comunicada; es decir, en una *cadena de Markov irreducible* es posible llegar de un estado a cualquier otro estado. Si la probabilidad de que nunca regresemos a un estado i, (estado de inicio), es diferente de cero, se dice que el estado i es un estado *transitorio*. Si el estado i no es transitorio, entonces se dice que es *recurrente*. Un estado i es *absorbente* si la probabilidad de salir de éste es cero, o sea, i es *absorbente* si y sólo si $p(i, i) = 1$ y $p(i, j) = 0$ para $i \neq j$.

4.1 Caminatas aleatorias

Una *caminata aleatoria simple* en un grafo G es una cadena de Markov finita donde cada vértice $v \in V$ corresponde a un estado y la probabilidad de transición del estado i al estado j es $p_{ij} = d_i^{-1}$ si $(i, j) \in E$ y cero en otro caso. Para un grafo ponderado, p_{ij} es la razón entre el peso de la arista (i, j) y la suma total de las aristas incidentes a i. La caminata se mueve de un vértice a su vértice vecino, es decir, dado un vértice semilla s_0 se selecciona un vértice s_1 vecino de s_0 uniformemente al azar y se mueve a s_1; luego se mueve uniformemente al azar a un vecino s_2 de s_1. Ésto se repite hasta que la caminata se ha movido un cierto número de *pasos*. Las caminatas que son realizadas de éste modo se conocen como *caminatas aleatorias ciegas*. Cada vez que la caminata se mueve a un vértice s_i, se dice que el vértice ha recibido una visita. La secuencia formada por los vértices visitados es conocida como un *camino*,

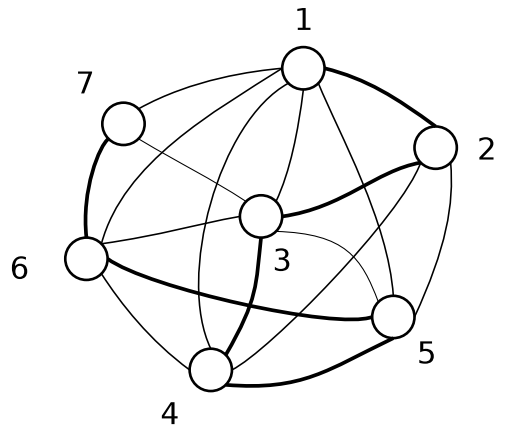


Figura 4.1: La secuencia de vértices {1,2,3,4,5,6,7} es un camino.

ver Figura 4.1.

Denotemos la matriz de probabilidad de transición de la cadena de Markov por $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$. Note que incluso para grafos no dirigidos, $\mathbf{P}$ no es necesariamente simétrica. Sin embargo la ecuación (4.3) es una matriz similar

$$\mathcal{P} = \mathbf{D}^{-\frac{1}{2}}\mathbf{P}\mathbf{D}^{-\frac{1}{2}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}, \tag{4.3}$$

la cual es simétrica ya que $\mathbf{A}$ es la matriz adyacente de un grafo no dirigido. Así, $\mathbf{P}$ y $\mathcal{P}$ tienen el mismo espectro de valores propios, los cuales son reales. La matriz $\mathcal{L}$ también puede ser expresada en términos de $\mathcal{P}$

$$\begin{aligned} \mathcal{L} &= \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{D}-\mathbf{A})\mathbf{D}^{-\frac{1}{2}} \\ &= \mathbf{D}^{-\frac{1}{2}}(\mathbf{D}-\mathbf{D}\mathbf{P})\mathbf{D}^{-\frac{1}{2}} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{P}\mathbf{D}^{-\frac{1}{2}} \\ &= \mathbf{I} - \mathcal{P}. \end{aligned} \tag{4.4}$$

Consecuentemente, λ es un valor propio de la matriz $\mathcal{P}$ si y sólo si $\mu = 1 - \lambda$ es un valor propio de la matriz Laplaciana normalizada $\mathcal{L}$. Así, $\mathbf{P}$, $\mathcal{P}$ y $\mathcal{L}$ tienen la siguiente correspondencia: $\mathbf{v}$ es un vector propio derecho asociado al valor propio λ en $\mathbf{P}$ si y sólo si $\mathbf{u} = \sqrt{\mathbf{D}}\mathbf{v}$ es un vector propio derecho asociado al mismo valor propio en $\mathcal{P}$, y para el valor propio $1 - \lambda$ en $\mathcal{L}$. Denotemos los valores propios de $\mathbf{P}$ en orden decreciente $\lambda_0^{(\mathbf{P})} \geq \lambda_1^{(\mathbf{P})} \geq \cdots \geq \lambda_{n-1}^{(\mathbf{P})}$. Dado que $\mathbf{P}$ es un matriz estócastica,

$\lambda_0^{(\mathbf{P})} = 1$, correspondiendo al valor propio más pequeño $\mu_0^{\mathcal{L}}$ de la matriz Laplaciana $\mathcal{L}$. El resto de los valores propios de $\mathbf{P}$ satisfacen $|\lambda_i^{(\mathbf{P})}| \leq 1$. Si además G es conexo y no bipartito, la cadena de Markov determinada por $\mathbf{P}$ es ergódica, en tal caso $|\lambda_i^{(\mathbf{P})}| < 1$ para todo $i \geq 1$. Sin mucha pérdida de generalidad, nosotros asumiremos está condición, además que todos los valores propios $\lambda_i^{(\mathbf{P})}$ son no negativos. Ambas condiciones pueden ser cumplidas considerando, si es necesario, en lugar de $\mathbf{P}$ la matriz de probabilidad de transición de la caminata aleatoria lenta (inglés: *lazy*)

$$\mathbf{P}' = \frac{1}{2}(\mathbf{I} + \mathbf{P}). \tag{4.5}$$

Para un grafo conexo G, está cadena es ergódica, y tiene valores propios no negativos

$$\lambda_{\mathbf{P}'}^{(i)} = \frac{1}{2}(1 + \lambda_{\mathbf{P}}^{(i)}), \tag{4.6}$$

con vectores propios iguales a los de $\mathbf{P}$. Entonces vamos a considerar una matriz de transición de probabilidad $\hat{\mathbf{P}}'$ obtenida de $\mathbf{P}'$ por hacer un estado s dado absorbente, el cual es el vértice semilla. Así, $\hat{\mathbf{P}}'$ es por lo anterior igual a $\mathbf{P}$, pero con todo $\hat{p_{si}} = 0$ excepto para $\hat{p_{ss}} = 1$. De aquí en adelante asumiremos, por simplicidad de notación, que $s = n$, así que en particular $\hat{\mathbf{P}}'$ tiene la siguiente estructura de bloques

$$\hat{\mathbf{P}}' = \left(\begin{array}{c|c} & p_1 \\ \mathbf{Q} & \vdots \\ & p_{n-1} \\ \hline 0 \cdots 0 & 1 \end{array}\right). \tag{4.7}$$

El *tiempo de absorción* m_i desde el vértice $i \neq s$ al vértice semilla s es el número esperado de pasos de una caminata que inició en i antes de llegar a s por primera vez. Intuitivamente, los tiempos de absorción miden en un cierto sentido la proximidad del vértice i al vértice s. Los vértices pertenecientes a un buen grupo S para s, si tal grupo existe, debe tener característicamente tiempos de absorción más pequeños que los vértices en $V \backslash S$. Sabemos que no todos los grafos presentan una estructura de grupos, en tal caso ningún método de agrupamiento será capaz de identificar un grupo de alta calidad [39].

Es bien conocido que los tiempos de absorción para el vértice $s = n$ pueden ser calculados a partir de la matriz fundamental $\mathbf{M}$

$$\mathbf{M} = \mathbf{I} + \mathbf{Q} + \mathbf{Q}^2 + \mathbf{Q}^3 + \ldots = (\mathbf{I} - \mathbf{Q})^{-1}, \tag{4.8}$$

donde $\mathbf{Q}$ es una matriz subestocástica obtenida de $\hat{\mathbf{P}}'$ (o equivalentemente de $\mathbf{P}'$) eliminando la fila y la columna correspondiente al vértice $s = n$ (como se muestra en la ecuación 4.7), como la suma por filas de $\mathbf{M}$

$$m_i = m_{i,1} + m_{i,2} + \ldots + m_{i,n-1}. \tag{4.9}$$

En la Figura 4.2, ilustramos los tiempos de absorción del grafo de la Figura 3.1: calculamos con Matlab los tiempos de absorción de todos los vértices a un vértice aislado v, repetimos el cálculo para cada vértice v, y formamos una matriz donde cada columna representa el vector de tiempo de absorción correspondiente a cada vértice semilla. Hemos ordenado estos vectores de forma tal que todos los vectores correspondientes a los vértices de una "cueva" se colocan antes de los de la próxima cueva, y así sucesivamente. La matriz es visualizada como una imagen en escala de grises poniendo en negro el pixel en el que el tiempo de absorción $m_i \leq 10.6$ (esto es, en la diagonal donde el tiempo de absorción es cero y en los tiempos de absorción más pequeños así calculados), en blanco el pixel en el que el tiempo de absorción es 319.6, el cual es el máximo, y discretizando los valores intermedios a los 254 tonos de la escala de grises. Las cuevas pueden ser vistas como bloques de 5×5 sobre la diagonal, para hacer trivial el agrupamiento la matriz tiene un poco de ruido.

Ahora considere los valores propios de las matrices $\hat{\mathbf{P}}$ y $\mathbf{Q}$. La matriz $\hat{\mathbf{P}}$ es estocástica, por lo que su mayor valor propio es $\lambda_0^{(\hat{\mathbf{P}})} = 1$, y dado que la cadena es absorbente, el resto de sus valores propios satisfacen que $|\lambda_i^{(\hat{P})}| < 1$, para $i = 1, \ldots, n-1$.

Denotemos a $\mathcal{Q} = \mathbf{D}^{-\frac{1}{2}}\mathbf{Q}\mathbf{D}^{-\frac{1}{2}}$, donde $\mathbf{D} = \text{diag}(d_1, \ldots, d_{n-1})$. Como $\mathcal{Q}$ es simétrica (ésta es obtenida de la matriz simétrica $\mathcal{P}$ eliminando la fila y la columna correspondientes al vértice semilla n) y $\mathbf{Q}$ es similar a $\mathcal{Q}$, ambas tienen espectro de

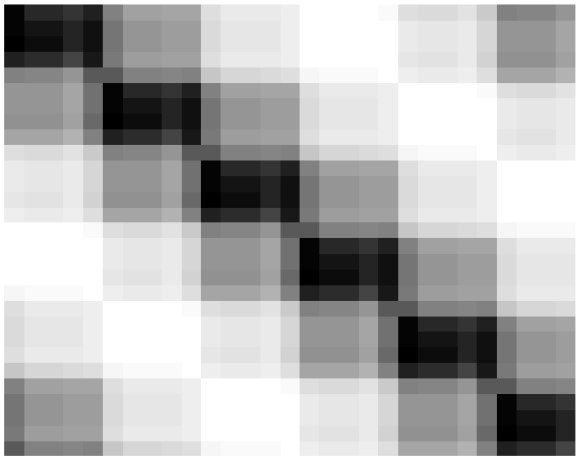


Figura 4.2: La matriz de tiempos de absorción compuesta por 30 vectores de tiempos de absorción usando cada vértice del grafo de la Figura 3.1 como un vértice semilla. En blanco se representa al m_{ij} máximo, en negro al m_{ij} mínimo y ceros en la diagonal.

valores propios reales $\text{Spec}(\mathbf{Q}) = \{\lambda_1^{(\mathbf{Q})} \geq \cdots \geq \lambda_{n-1}^{(\mathbf{Q})}\}$. Este espectro está propiamente contenido en el intervalo $[-1, 1]$, porque para algún vértice $i \neq n$ adyacente a n, $p_{in} > 0$, así la suma de la $i-$ésima fila de $\mathbf{Q}$ es menor que 1.

Aplica que [36]:

$$\text{Spec}(\mathbf{Q}) = \text{Spec}(\hat{\mathbf{P}}) \backslash \{1\}. \tag{4.10}$$

Para probar está afirmación, sea $\lambda \neq 1$ algún valor propio no principal de $\hat{\mathbf{P}}$ y sea $\mathbf{v}$ el vector propio correspondiente tal que $\hat{\mathbf{P}}\mathbf{v} = \lambda\mathbf{v}$. Dado que la n $-$ésima fila de $\hat{\mathbf{P}}$ es cero excepto para $\hat{p}_{nn} = 1$, de ahí se deduce que $\lambda v_n = (\mathbf{P}\mathbf{v})_n = v_n$ y de $\lambda \neq 1$ necesariamente $v_n = 0$. Entonces para el vector de dimensión $(n-1)$ $\mathbf{v}' = (v_1, \ldots, v_{n-1})$ y para algún $i = 1, \ldots, n-1$ tenemos que:

$$\begin{aligned} (\mathbf{Q}\mathbf{v}')_i &= \sum_{j=1}^{n-1} p_{ij} v'_j = \sum_{j=1}^{n-1} p_{ij} v_j = \sum_{j=1}^{n} p_{ij} v_j - p_{in} v_n \\ &= (\hat{\mathbf{P}}\mathbf{v})_i - v_n p_{in} = (\hat{\mathbf{P}}\mathbf{v})_i \\ &= \lambda v_i = \lambda v'_i. \end{aligned} \tag{4.11}$$

Consecuentemente, $\mathbf{v}'$ es un vector propio asociado al valor propio λ de $\mathbf{Q}$. Dado que λ fue elegido arbitrariamente de $\text{Spec}(\hat{\mathbf{P}}) \backslash \{1\}$, esto establece que $\text{Spec}(\hat{\mathbf{P}}) \backslash \{1\} \subseteq \text{Spec}(\mathbf{Q})$. Para el recíproco se prueba un argumento similar, si $\mathbf{v}' = (v_1, \ldots, v_{n-1})$ es

un vector propio asociado al valor propio λ de $\mathbf{Q}$, entonces el vector $\mathbf{v} = (v_1, \ldots, v_{n-1}, 0)$ es un vector propio asociado al valor propio λ de $\hat{\mathbf{P}}$.

4.2 Partición espectral como un problema de relajación de un programa entero

El problema de partición espectral puede ser modelado como un problema de optimización [27]. El uso del vector de Fiedler para biparticionar grafos puede ser motivado de la siguiente manera (ver por ejemplo [20, 23]). Denote un *corte* (bipartición) de un grafo $G = (V, E)$ por los conjuntos de vértices S y $\bar{S} = V \backslash S$ como $(S, \bar{S})$. La *capacidad* de un corte $(S, \bar{S})$ es definida como

$$C(S, \bar{S}) = \left|\{\{i, j\} \in E : i \in S, j \in \bar{S}\}\right|. \tag{4.12}$$

Un corte $(S, \bar{S})$ puede ser convenientemente representado por un vector indicador $\mathbf{v} \in \{+1, -1\}^n$, donde $v_i = +1$ si $i \in S$ y $v_i = -1$ si $i \in \bar{S}$. Entonces

$$C(S, \bar{S}) = \frac{1}{4} \sum_{i \sim j} (v_i - v_j)^2, \tag{4.13}$$

donde la suma es sobre todos los aristas (no dirigidos) $(i, j) \in E$.

Por simplicidad, asumimos ahora que $|V| = n$ es par, y consideramos que la tarea es encontrar una bisección óptima de G, es decir, un corte $(S, \bar{S})$ que satisfaga la condición $|S| = |\bar{S}| = n/2$ y que minimice $C(S, \bar{S})$ sujeto a esta condición.

Lo anterior es equivalente a encontrar un vector indicador $\mathbf{v} \in \{+1, -1\}^n$ que satisfaga que $\sum_i v_i = 0$ y minimice la forma cuadrática $\sum_{i \sim j} (v_i - v_j)^2$, o equivalentemente (dado un n fijo) minimice la razón:

$$\begin{aligned} \frac{\frac{1}{4} \sum_{i \sim j} (v_i - v_j)^2}{n/4} &= \frac{\sum_{i \sim j} (v_i - v_j)^2}{n} \\ &= \frac{\sum_{i \sim j} (v_i - v_j)^2}{\sum_i v_i^2}. \end{aligned}$$

Dado que el vector $\mathbf{1}$ de unos está asociado al valor propio $\mu^{\mathbf{L}}_{(0)} = 0$ y por la caracterización de Courant-Fisher del valor propio diferente de cero más pequeño $\mu^{\mathbf{L}}_{(1)}$ tenemos que:

$$\mu^{1}_{(\mathbf{L})} \quad = \quad \min_{\mathbf{v}\perp\mathbf{1}} \frac{\mathbf{v}^{\mathrm{T}} L\mathbf{v}}{\mathbf{v}^{\mathrm{T}}\mathbf{v}} \quad = \quad \min_{\sum_i v_i = 0} \frac{\sum_{i\sim j}(v_i - v_j)^2}{\sum_i v_i^2}, \tag{4.14}$$

donde el mínimo se toma sobre todos los vectores $v \neq 0$ que satisfacen la condición dada. Dado que nosotros también podemos, sin pérdida de generalidad, restringir la minimización, es decir, la norma de los vectores $\|\mathbf{v}\|^2 = n$, vemos que la tarea de encontrar un vector de Fiedler de G es en realidad una *relajación fraccional* del problema combinatorio de determinar una bisección óptima de G.

Está correspondencia motiva la aproximación espectral indicada previamente a particionar un grafo conexo G [16, 18] como:

1. Calcular el vector de Fiedler $\mathbf{v} \in \mathbb{R}^n$ de G.

2. Determinar el corte $(S, \bar{S})$ por:

$$\begin{cases} v_i \geq 0 \;\; \Rightarrow \;\; i \in S, \\ v_i < 0 \;\; \Rightarrow \;\; i \in \bar{S}. \end{cases} \tag{4.15}$$

El uso del vector de Fiedler *normalizado* para biparticionar grafos fue explorado anteriormente [40], y se demostró que el vector de Fiedler de $\mathcal{L}$ da biparticiones óptimas de acuerdo a la medida de la *capacidad del corte normalizado*:

$$\hat{C}(S, \bar{S}) = \frac{C(S, \bar{S})}{\mathrm{Vol}(S)} + \frac{C(S, \bar{S})}{\mathrm{Vol}(\bar{S})}, \tag{4.16}$$

donde $\mathrm{Vol}(S) = \sum_{i\in S} d_i$. Dado que $\mathbf{u}$ es un vector propio de $\mathcal{L}$ con valor propio λ si y sólo si $\mathbf{v} = \mathbf{D}^{-\frac{1}{2}}\mathbf{u}$ es vector propio de $\mathbf{D}^{-1}\mathbf{L}$ con valor propio λ, el valor propio $\mu^{\mathcal{L}}_{(1)}$ puede ser caracterizado en terminos de "ajustar el grado" con el *cociente de Rayleigh*:

$$\mu^{\mathcal{L}}_{(1)} = \min_{\mathbf{u}\perp\sqrt{\mathbf{D}}\mathbf{1}} \frac{\mathbf{u}^{\mathrm{T}}\mathcal{L}\mathbf{u}}{\mathbf{u}^{\mathrm{T}}\mathbf{u}} = \min_{\mathbf{v}\perp\mathbf{D}\mathbf{1}} \frac{\sum_{i\sim j}(v_i - v_j)^2}{\sum_i d_i v_i^2}. \tag{4.17}$$

Una extensión natural del agrupamiento espectral al agrupamiento local es la idea de considerar la matriz Laplaciana $\mathbf{L}$ o $\mathcal{L}$ junto con la *condición de frontera de*

Dirichlet y la idea de que los vectores $\mathbf{v}$ con el vértice semilla s fijo pueden dar soluciones aceptables. Usando la matriz Laplaciana normalizada $\mathcal{L}$ y eligiendo $v_s = 0$, o equivalentemente $u_s = (\sqrt{\mathbf{D}}\mathbf{v})_s = 0$ como la condición de frontera [10, 12]. Nuestro objetivo de acuerdo al grupo para el "vector de Dirichlet-Fiedler" es minimizar la restricción del cociente de Rayleigh:

$$\min_{\mathbf{u}:u_s=0} \frac{\mathbf{u}^{\mathrm{T}}\mathcal{L}\mathbf{u}}{\mathbf{u}^{\mathrm{T}}\mathbf{u}} = \min_{\mathbf{v}:v_s=0} \frac{\sum_{i\sim j}(v_i - v_j)^2}{\sum_i d_i v_i^2}. \tag{4.18}$$

Por simplicidad de notación, asumimos nuevamente que $s = n$, observe que para cada vector $\mathbf{u} = (u_1, \ldots, \mathbf{u}_{n-1}, 0)$, el valor del cociente de Reyleigh en la ecuación 4.18 es el mismo respecto al vector $\mathbf{u}' = (u_1, \ldots, \mathbf{u}_{n-1})$ y $\mathcal{L}'$, la cual es igual a $\mathcal{L}$ con la n–ésima fila y columna removidas. Por tanto, nuestro vector $\mathbf{v}$ para el agrupamiento es, excepto para el cero final, la minimización de:

$$\min_{\mathbf{u}'} \frac{(\mathbf{u}')^{\mathrm{T}}\mathcal{L}'\mathbf{u}'}{(\mathbf{u}')^{\mathrm{T}}\mathbf{u}'} = \min_{\mathbf{v}'} \frac{\sum_{i\sim j}(v'_i - v'_j)^2}{\sum_i d_i (v'_i)^2}, \tag{4.19}$$

es decir, $\mathbf{v}' = \mathbf{D}^{-\frac{1}{2}}\mathbf{u}'$ para el vector propio principal $\mathbf{u}'$ de la matriz Laplaciana $\mathcal{L}'$. Denotemos $\mathbf{v} = \mathbf{v}^f$, y sea este el *vector local de Fiedler* asociado al grafo G y el vértice semilla $s = n$.

Capítulo 5

Vector de Fiedler local y tiempos de absorción

Ahora probaremos que las componentes del vector de Fiedler $\mathbf{v}^f = (v_1, \ldots, v_{n-1})$ son aproximadamente proporcionales a los tiempos de absorción m_i descritos en la sección 4.1. La conexión entre los tiempos de absorción proveen una interpretación natural de la noción del vector local de Fiedler, y da más soporte a la idea del agrupamiento local restringiendo las técnicas espectrales. Anteriormente las caminatas aleatorias y el agrupamiento espectral han sido dirigidos por Meila y Shi [31] y el agrupamiento local para PageRank por Andersen, Chung y Lang [2]. Existen propiedades espactrales de grafos ligadas a tasas de convergencia de caminatas aleatorias [42]. Observemos primero que de la ecuación (4.5):

$$\mathcal{L}' = \mathbf{I} - \sqrt{\mathbf{D}}\mathbf{Q}\mathbf{D}^{-\frac{1}{2}} = \mathbf{I} - \mathcal{Q}, \tag{5.1}$$

donde $\mathbf{Q}$ es la matriz fundamental (cf. la sección 4.1) y $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_{n-1})$. Dado que $\mathcal{Q}$ es similar a $\mathbf{Q}$, su espectro satisface que:

$$\mathrm{Spec}(\mathcal{Q}) = \mathrm{Spec}(\mathbf{Q}) = \mathrm{Spec}(\hat{\mathbf{P}})\backslash\{1\}. \tag{5.2}$$

Así, $\mu \neq 0$ es un valor propio de $\mathcal{L}'$ si y sólo si $\lambda = 1 - \mu \neq 1$ es un valor propio de $\mathcal{Q}$ y $\mathbf{Q}$. Además, si $\mathbf{u}$ es un vector propio asociado al valor propio λ en $\mathcal{Q}$, entonces $\mathbf{v} = \mathbf{D}^{-\frac{1}{2}}\mathbf{u}$ es un vector propio asociado al mismo valor propio en $\mathbf{Q}$.

Sean entonces los valores propios de $\mathcal{Q}$ (igualmente $\mathbf{Q}$) $1 > \lambda_1 \geq \cdots \geq \lambda_{n-1}$. Dado que $\mathcal{Q}$ es simétrica, ésta tiene un sistema ortonormal de vectores propios

$\mathbf{u}_1, \ldots, \mathbf{u}_{n-1}$ y una representación [36]:

$$\mathcal{Q} = \sum_{i=1}^{n-1} \lambda_i \mathbf{u}_i \mathbf{u}_i{}^{\mathrm{T}}. \tag{5.3}$$

Denotemos por $\mathbf{U}$ la matriz con componentes $\mathbf{U}_i = \mathbf{u}_i \mathbf{u}_i{}^{\mathrm{T}}$, observemos ahora que por ortogonalidad de los vectores propios tenemos que $\mathbf{U}_i \mathbf{U}_j = 0$ para $i \neq j$, y por normalidad $\mathbf{U}_i^2 = \mathbf{U}_i$. De las dos observaciones anteriores se sigue que:

$$\mathcal{Q}^t = \sum_{i=1}^{n-1} \lambda_i^t \mathbf{U}_i, \qquad \text{para } t = 0, 1, \ldots \tag{5.4}$$

Dado que $\mathbf{Q} = \mathbf{D}^{-\frac{1}{2}} \mathcal{Q} \sqrt{\mathbf{D}}$, podemos obtener una representación para $\mathbf{Q}^T$:

$$\mathbf{Q}^t = \mathbf{D}^{-\frac{1}{2}} \mathcal{Q}^t \sqrt{\mathbf{D}} = \sum_{i=1}^{n-1} \lambda_i^t (\mathbf{D}^{-\frac{1}{2}} \mathbf{u}_i)(\mathbf{u}_i{}^{\mathrm{T}} \sqrt{\mathbf{D}}) = \sum_{i=1}^{n-1} \lambda_i^t \mathbf{v}_i \mathbf{v}_i{}^{\mathrm{T}} \mathbf{D}, \tag{5.5}$$

donde $\mathbf{v}_i = \mathbf{D}^{-\frac{1}{2}} \mathbf{u}_i$ es un vector propio asociado al valor propio λ_i de $\mathbf{Q}$.

Sustituyendo ésto en la ecuación 4.8 y denotando el vector $\mathbf{1}$ de dimensión $(n-1)$, así obtenemos una expresión para el vector $\mathbf{m}$ de los tiempos de absorción m_i en términos de los valores y vectores propios de $\mathbf{Q}$, o equivalentemente $\mathcal{Q}$:

$$\begin{aligned} \mathbf{m} &= \sum_{t=0}^{\infty} \mathbf{Q}^t \mathbf{1} \\ &= \sum_{t=0}^{\infty} \sum_{i=1}^{n-1} \lambda_i^t \mathbf{v}_i \mathbf{v}_i{}^{\mathrm{T}} \mathbf{D} \mathbf{1} \\ &= \sum_{t=0}^{\infty} \left(\sum_{i=1}^{n-1} \lambda_i^t \mathbf{v}_i \mathbf{v}_i{}^{\mathrm{T}} \right) \mathbf{d}, \end{aligned} \tag{5.6}$$

donde $\mathbf{d} = (d_1, \ldots, d_{n-1})^{\mathrm{T}}$.

Ahora, si el valor propio λ_1 está *bien separado* del resto, es decir, si la razón $|\lambda_i / \lambda_1|$ es pequeño para $i < 1$, esto nos da una buena aproximación para $\mathbf{m}$:

$$\begin{aligned} \mathbf{m} &= \mathbf{1} + \sum_{t=1}^{\infty} \lambda_1^t \left(\mathbf{v}_1 \mathbf{v}_1{}^{\mathrm{T}} \mathbf{d} + \underbrace{\sum_{i=2}^{n-1} \left(\frac{\lambda_i}{\lambda_1} \right)^t \mathbf{v}_i \mathbf{v}_i{}^{\mathrm{T}}}_{\text{posiblemente tiende a cero}} \right) \\ &\approx \mathbf{1} + \sum_{t=1}^{\infty} \lambda_1^t \mathbf{v}_1 \mathbf{v}_1{}^{\mathrm{T}} \mathbf{d} \\ &= \mathbf{1} + \frac{\lambda_1}{1 - \lambda_1} \mathbf{v}_1 \mathbf{v}_1{}^{\mathrm{T}} \mathbf{d}. \end{aligned} \tag{5.7}$$

Incluso en casos donde no existe una diferencia (inglés: *gap*) evidente en el espectro no podemos asumir nada acerca de la igualdad, pero hemos encontrado en los experimentos (capítulo 6) que la aproximación obtenida está *casi perfectamente* correlacionada con los tiempos de absorción exactos en los diferentes grafos de experimentación.

En la práctica, no siempre es de interés calcular los tiempos de absorción para todos los vértices, especialmente en el cálculo local. En tal caso, nosotros podemos aproximar algunas de las componentes del vector de Fiedler. Para éste caso, podemos escribir la $k-$ésima componente del vector como:

$$\begin{aligned}(\mathbf{Q}^t\mathbf{1})_k &= \left(\sum_{i=1}^{n-1}\lambda_i^t\mathbf{v}_i{\mathbf{v}_i}^{\mathrm{T}}\mathbf{D}\mathbf{1}\right)_k \\ &= \left(\sum_{i=1}^{n-1}\lambda_i^t(\mathbf{v}_i{}^{\mathrm{T}}\mathbf{d})\mathbf{v}_i\right)_k \\ &= \sum_{i=1}^{n-1}\lambda_i^t(\mathbf{v}_i)_k\underbrace{\sum_{\ell=1}^{n-1}(\mathbf{v}_i)_\ell(\mathbf{d})_\ell}_{c_i}.\end{aligned} \tag{5.8}$$

De la ecuación 5.8 podemos obtener una expresión para el tiempo de absorción de un vértice k al vértice s:

$$\begin{aligned}m_k &= \sum_{t=0}^{\infty}(\mathbf{Q}^t\mathbf{1})_k \\ &= 1+\sum_{t=1}^{\infty}\lambda_1^t\left(c_i\cdot(\mathbf{v}_i)_k+\sum_{i=2}^{n-1}\left(\frac{\lambda_i}{\lambda_1}\right)^t c_i\cdot(\mathbf{v}_i)_k\right) \\ &\approx 1+\sum_{t=1}^{\infty}\lambda_1\cdot c_1\cdot(\mathbf{v}_1)_k \\ &= 1+\underbrace{\frac{\lambda_1}{1-\lambda_1}\cdot c_1}_{c'}\cdot(\mathbf{v}_1)_k.\end{aligned} \tag{5.9}$$

Ahora para un grafo G dado, c' es una constante y por lo tanto obtenemos la aproximación correspondiente $\mathbf{m}\approx\mathbf{1}+c'\mathbf{v}^f$ entre el vector de tiempos de absorción $\mathbf{m}$ y el vector local de Fiedler $\mathbf{v}^f=\mathbf{v}_1$.

5.1 Aproximación local del vector de Fiedler

Para poder aproximar localmente el vector de Fiedler se puede tomar como punto de inicio el cociente de Rayleigh de la ecuación (4.19) [35]. Dado que se está normalizando libremente el vector de Fiedler eventual v^f por alguna longitud deseada, se puede restringir la minimización a vectores que cumplan $\|\mathbf{v}\|_2^2 = n = |V|$. Así, la tarea es encontrar un vector $\mathbf{v}$ para un vértice semilla $s \in V$ que satisfaga:

$$\mathbf{v}^f \quad = \quad \text{argmin}\left\{ \sum_{j\sim k}(v_j - v_k)^2 : v_s = 0,\ \|\mathbf{v}\|_2^2 = n \right\}. \tag{5.10}$$

Está tarea se puede resolver aproximadamente reformulando la condición $\|\mathbf{v}\|_2^2 = n$ como una "restricción suave" con un peso $c > 0$, y minimizando la función objetivo [35]:

$$f(\mathbf{v}) \quad = \quad \frac{1}{2}\sum_{j\sim k}\left(v_j - v_k\right)^2 + \frac{c}{2}\cdot\left(n - \sum_j v_j^2\right) \tag{5.11}$$

por el método del gradiente descendente. Dado que las derivadas parciales de f tienen la forma:

$$\frac{\partial f}{\partial v_j} \quad = \quad -\sum_{k\sim j} v_k + (\deg j - c)\cdot v_j, \tag{5.12}$$

el paso de descenso puede ser calculado localmente para cada vértice en el tiempo $t+1$.

Es posible que la solución converja a la solúción exacta incluso en grafos dirigidos en donde sólo conocemos las aristas que salen del vértice de interés, ésto porque localmente el vértice semilla solo puede obtener información de los vértices con que él se comunica. No puede obtener información de los vértices que entran a él y ésto hace que la propagación de información para saber si vértices que entran a él pertencean a su grupo sea lenta. Lo anterior porque el vértice semilla tendría que enviar información a alguno de los vértices que salen de él y este a su vez buscar por algún camino que lo comunique con el vértice que entra al vértice semilla.

Para grafos no dirigidos el paso de descenso es más rápido dado que las arista no tienen dirección. Ésto hace que la solución converja más rápidamente. Se espera

que la convergencia en el caso de grafos dirigidos sea más lenta y dejamos como trabajo futuro la comprobación experimental de tal hipótesis. Por ahora continuaremos refiriendonos a grafos simples.

Basandonos en la información acerca de los valores del vector $\mathbf{v}$ en el tiempo t, denotamos por $\tilde{\mathbf{v}}(t)$ al vértice mismo y sus vecinos:

$$\tilde{v}_j(t+1) \quad = \quad \tilde{v}_j(t) + \delta \cdot \left(\sum_{k \sim j} \tilde{v}_k - (\deg j - c) \cdot \tilde{v}_j \right), \tag{5.13}$$

donde $\delta > 0$ es un parámetro que determina la velocidad del descenso.

Asumiendo que el grupo natural del vértice s es pequeño comparado con el orden n del grafo, la normalización $\|\mathbf{v}\|_2^2 = n$ implica que la mayoría de los vértices j en el grafo deberían tener $v_j \approx 1$. De este modo, las iteraciones de descenso (5.13) se pueden empezar de un vector inicial $\tilde{\mathbf{v}}(0)$ que tenga $\tilde{v}_s(0)$ para el vértice semilla $s \in V$ y $\tilde{v}_k(0) = 1$ para todo $k \neq i$. Las estimaciones necesitan ser actualizadas en el tiempo $t > 0$ sólo para aquellos vectores j que tienen al menos un vecino k tal que $\tilde{v}_k(t-1) < 1$.

Balancear la restricción del peso c en contra de la velocidad del gradiente descendente δ requiere algo de atención. Se han obtenido [35] razonablemente resultados estables con la siguiente heurística: dado un valor estimado $\overline{k}$ para el grado promedio de los vértices en la red, el conjunto $c = 1/\overline{k}$ y $\delta = c$, las iteraciones del gradiente 5.13 se calculan hasta que todos los cambios en las aproximaciones de v estén por debajo de $\epsilon = \delta/10$.

Los valores aproximados del vector de Fiedler obtenidos representan valores de proximidad entre los vértices de V para el grupo de vértices de s. Determinar una bisección en S y $V \backslash S$ es una tarea de dos-clasificación unidimensional que en principio se puede resolver usando alguno de los patrones de clasificación estándar, tales como las variaciones del algoritmo de las $k-$medias [22].

Capítulo 6

Agrupamiento por caminatas aleatorias

En esté capítulo mostramos los resultados obtenidos para el agrupamiento con la aproximación del vector de Fiedler a través de los tiempos de absorción y el agrupamiento con caminatas aleatorias.

Recordemos de la sección 4.1 que una caminata aleatoria es un proceso en el que dado un grafo G y un vértice s_0 de inicio "semilla" nos movemos uniformemente al azar a un vecino s_1 de s_0, luego nos movemos nuevamente uniformemente al azar a un vértice vecino s_3 de s_2, repetimos sucesivamente este proceso hasta que la caminata cumpla un determinado número de pasos.

El cálculo de la matriz de tiempos de absorción $\mathbf{M}$, ecuación 4.8, es muy costoso computacionalmente. Este cálculo es una operación global lo cual implica que su complejidad es peor que la cuadrática [13] por lo que las operaciones con matrices de grandes dimensiones se complican y en ocasiones es imposible trabajar con toda la matriz. Por esta razón se propone realizar la aproximación del agrupamiento local por medio de tiempos de absorción con caminatas aleatorias *cortas* sobre un grafo dado G. Es importante que el número de pasos de las caminatas sea pequeño para que la caminata no alcance su equilibrio. Cuando una cadena ergódica está en equilibrio ya no se puede observar estadísticamente en cuál vértice inició la caminata: su distribución estacionaria depende unicamente del grado de los vértices [29].

Recordemos que en el caso de agrupamiento local el interés es conocer los vértices cercanos a un vértice semilla s y no propiedades globales, por lo cual la utilidad de las *frecuencias de visitas* f de las caminatas aleatorias largas se pierde junto con la información del vértice de inicio. La caminata aleatoria inicia en un vértice semilla y el número total de visitas que recibe cada vértice es llamado frecuencia de una visita, el número total de visitas es llamado frecuencia de visitas.

Para el cálculo de las caminatas aleatorias se ha desarrollado un algoritmo [4] el cual describiremos ahora. La entrada del algoritmo es un listado de las aristas del grafo, indicando en la primera línea el número total de vértices y aristas, junto con el vértice de inicio. Se utilizan listas simplemente ligadas lineales para almacenar el grafo en forma de listas ordenadas de adyacencia. También se utilizan para almacenar las frecuencias de visitas a los vértices en V al repetir r caminatas de k pasos iniciadas en el vértice s. Los grados de los vértices son almacenados en un vector d. La frecuencia de visitas a un vértice v se guarda a partir de la primera visita en ello, por lo cual el tamaño del grafo de entrada no afecta directamente la memoria necesaria para almacenar las frecuencias.

Estudiamos tres ejemplos de grafos señalando las debilidades y fortalezas de nuestra aproximación. El primer grafo de la Figura 3.1 está agrupado de forma altamente simétrica y es llamado grafo de hombre de cueva, donde la simetría presente causa problemas para la aproximación propuesta. El segundo ejemplo es el grafo del club de karate mostrado en la Figura 3.2. El tercer ejemplo es un grafo $\mathcal{G}_{n,p}$ que generamos uniforme al azar, con $n = 100$ y $p = 0.1$ [21], el cual por definición no tiene una estructura de grupos clara y no podemos esperar que los tiempos de absorción tengan patrones evidentemente claros.

En la Figura 6.1, mostramos las comparaciones entre los tiempos de absorción exactos y los aproximados (ecuación 4.9 y ecuación 5.7) para los tres ejemplos, usando sólo el último vértice como vértice semilla. Cabe señalar que el espectro de estos grafos carece de una diferencia inicial grande, por lo cual no podemos esperar que los valores aproximados tengan inclusive la misma magnitud que los valores

exactos si sólo el primer valor y vector propio son considerados. Además, la simetría del grafo de hombre de cuevas es notable en el espectro: la diferencia se encuentra después de seis valores propios (uno por grupo).

La correlación es una medida sobre el grado de relación entre dos variables, sin importar cual es la causa y cual es el efecto. La dependencia de la que se habla en este sentido es la dependencia entre la varianza de las variables: graficamente, cuando existe una correlación alta, observamos una dependencia lineal. En los tres grafos estudiados la correlación es alta: para el grafo de hombre de cuevas obtenemos una correlación de 0.99863, para el grafo que representa el club de karate tenemos 0.99636, y en el caso del grafo uniformemente al azar $\mathcal{G}_{n,p}$ [21, 45] observamos una correlación de 0.99999. También graficamos el vector exacto y el aproximado. De hecho, en el grafo de hombre de cuevas la estructura de los seis grupos es muy visible (Figura 3.1).

En la red del club de karate podemos ver que los dos grupos están presentes: uno con valores altos y otro con valores bajos como se muestra en la Figura 6.1. Como era de esperarse, el grafo aleatorio uniforme no muestra una estructura de grupos claros, pero los vértices muy cercanos al vértice semilla pueden ser identificados por sus valores pequeños, mientras que la mayor parte del grafo tiene valores grandes.

Con el fin de comparar la aproximación que proponemos, así como para ilustrar el cálculo computacional en la aproximación de la ecuación 5.6 sumamos término por término parcialmente

$$\begin{aligned}
\mathbf{M}_0 &= \mathbf{I} \\
\mathbf{M}_1 &= \mathbf{I} + \mathbf{Q} \\
\mathbf{M}_2 &= \mathbf{I} + \mathbf{Q} + \mathbf{Q}^2 \\
\mathbf{M}_3 &= \mathbf{I} + \mathbf{Q} + \mathbf{Q}^2 + \mathbf{Q}^3 \\
&\vdots
\end{aligned}$$

hasta un número fijo de iteraciones. Calculamos en cada iteración la suma de cuadrados de la diferencia entre las sumas parciales y los tiempos de absorción exactos,

dividido por el orden de cada uno de los tres ejemplos: el grafo de la Figura 3.1, el grafo del club de karate de Zachary de la Figura 3.2 y el grafo aleatorio uniforme $\mathcal{G}_{n,p}$. Los resultados sobre los vértices se muestran en la Figura 6.2 (a la izquierda) junto con la correlación de Pearson (en el lado derecho) obtenidos en cada iteración. En ambas gráficas se muestra la media y la desviación estandar.

Para la Figura 3.2 se ilustran los tiempos de absorción para el agrupamiento de la red del club de karate. Los tiempos de absorción aproximados que son calcula-

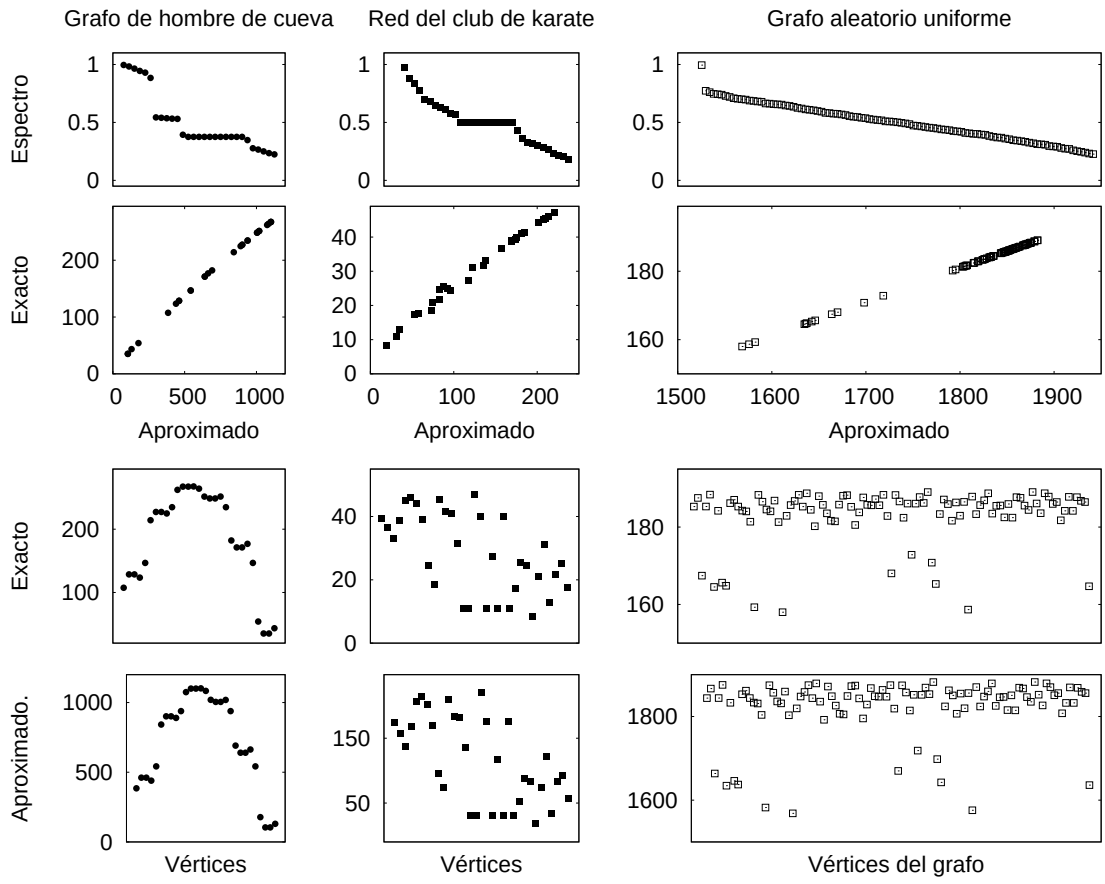


Figura 6.1: Comparación de los valores aproximados y exactos (ecuación 4.9) de los tres ejemplos de grafos: el de hombre de cueva de la Figura 3.1, el grafo del club de karate Figura 3.2 y el grafo $\mathcal{G}_{n,p}$, usando un vértice aleatorio como vértice semilla. En la primera fila mostramos el espectro ordenado de cada grafo, en la segunda la correlación entre el valor del vector exacto y el aproximado, y en la tercer y cuarta fila mostramos los valores del vector exacto y aproximado.

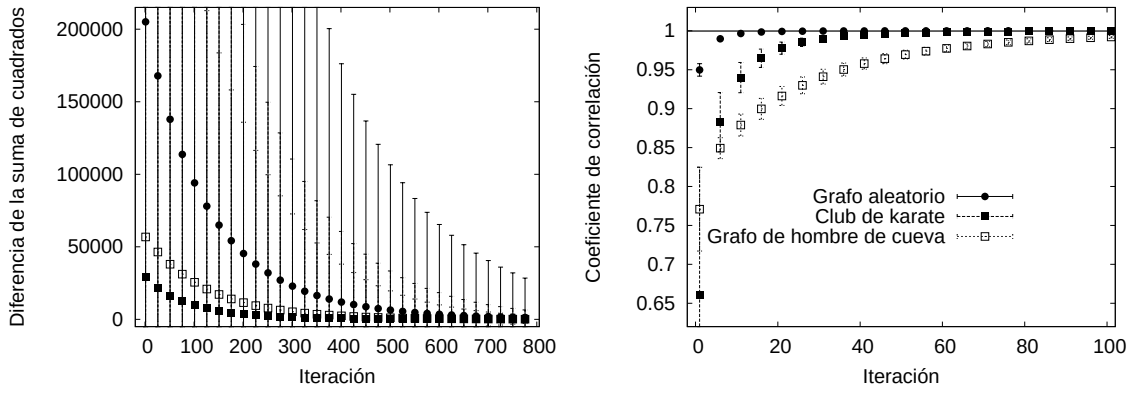


Figura 6.2: La suma de cuadrados exacta de la diferencia de los tiempos de absorción (a la izquierda) de los vectores estimados con diferentes valores para la aproximación a través de la ecuación 5.6 y la correlación de Pearson entre los valores de los vectores exactos y aproximados (a la derecha para los tres ejemplos de grafos: el grafo de las Figuras 3.1 y 3.2), y el grafo $\mathcal{G}_{n,p}$. Los valores mostrados son las medias sobre conjuntos de vértices de los dos primeros grafos, y sobre un conjunto de 100 vértices seleccionados uniformemente al azar para el grafo $\mathcal{G}_{n,p}$. La desviación estándar más pequeña corresponde al grafo de hombre de cuevas y la más grande al grafo aleatorio uniforme. Las líneas horizontales (las tres se traslapan entre 0.980 y 0.997) corresponden a las medias de los coeficientes de correlación entre los tiempos de absorción exactos y aproximados de la ecuación 5.7.

dos con la ecuación 5.9 se muestran en la Figura 6.3: observamos que la estructura del grupo es fuerte cuando el vértice semilla es uno de los miembros centrales del grupo, mientras que la tarea de clasificación es difícil para vértices que tienen relativamente poca comunicación con los vértices del grupo al que pertenecen, lo cual era de esperarse. Los agrupamientos de la Figura 3.5 página 20 fueron realizados con la aproximación local del vector de Fiedler [35].

6.1 Experimentos de caminatas aleatorias

Para analizar la supuesta comparación entre los tiempos de absorción y las frecuencias de visita, se realizaron experimentos computacionales con un grafo no dirigido y un grafo dirigido, ambos con pocos vértices para poder hacer el cálculo

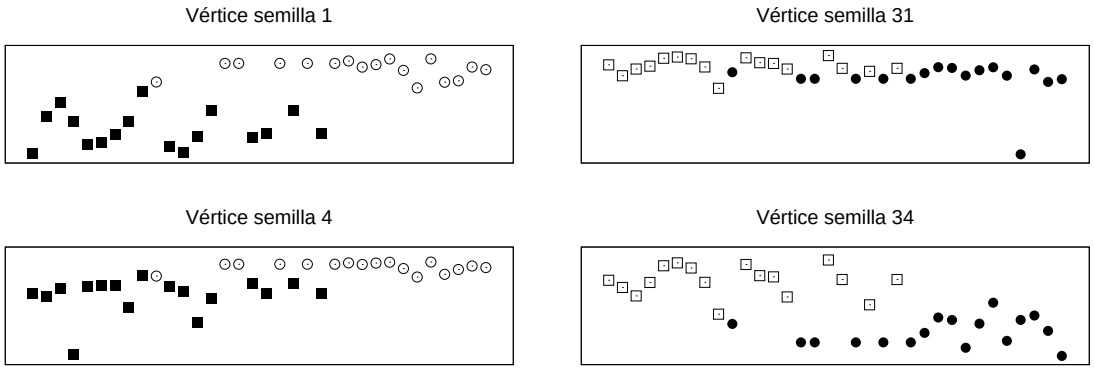


Figura 6.3: Cuatro ejemplos de dos-clasificación de vértices del la red del club de karate. Los ejemplos de la columna izquierda tienen vértice semilla representado por los rectángulos de la Figura 3.2 y los ejemplos de la columna derecha tiene vértice semilla representado por los círculos. Los vértices son ordenados en el mismo orden que muestra su etiqueta en la Figura 3.2 de la página 16 y en la posición correspondiente al vértice semilla se ha insertado un cero para representar absorción instantánea. El grupo al cual pertenece al vértice semilla está dibujado en color negro y el otro grupo en blanco.

exacto de la matriz $\mathbf{M}$ para cada vértice semilla. El primer grafo con el cual se realizó la comparación es el grafo de la Figura 3.1.

Para cada uno de los vértices se calculó su tiempo de absorción para determinar el grupo al cual pertenece cada uno de los vértices, en este caso en la Figura 6.4 mostramos ejemplos de la comparación de los grupos que se determinan por las frecuencias de visita y los tiempos de absorción, con ésto hemos comprobado la hipótesis de que mientras un vértice tenga tiempo de absorción pequeño comparado con el resto, pertenece al grupo del vértice semilla, vértices con frecuencia de visitas alta pertenecen al grupo del vértice de semilla.

Para determinar por medio de caminatas aleatorias los grupos, se calcularon las frecuencias usando $r = 1, 2, 3, \ldots, 30$ utilizando como vértice semilla s cada uno de los vértices. El proceso se repitió 30 veces por cada s para observar la variabilidad en los resultados. Para determinar el número de pasos k de la caminata, se buscó limitarnos al procesamiento *local* del grafo, por lo cual es necesario que k sea

mucho menor que el *diámetro*: si k es igual al diámetro, la probabilidad de que cada caminata visite a cualquier vértice es mayor que cero, por lo que la computación para determinar los grupos resulta ser global en un cierto sentido y no es ésto lo que queremos.

Hay grafos que no presentan una estructura de grupos naturales presentes. En éste caso ningún valor de k dará un agrupamiento claro para s y el resto del grafo. Para los grafos que si cuentan con tal estructura, el diámetro de un grupo es típicamente menor al diámetro del grafo. Lo que nosotros buscamos es un valor de k parecido al diámetro del grupo, lo que convierte a k en el parámetro más importante del método. Para valores demasiado pequeños, casi todos los vértices que son visitados tienen frecuencias altas debido a que las caminatas raramente salen del grupo, lo anterior hace posible que alguna parte del grupo no sea detectada. En cambio, cuando los valores de k son demasiado altos, las caminatas empiezan a visitar vértices con grado alto aunque no pertenezcan al grupo y por está razón dichos vértices son mal clasificados como miembros del grupo.

En la Figura 6.4 se muestran los tiempos de absorción desde el vértice $i = 1, 2, 3, \ldots$ junto con las frecuencias de visita con $k = 2$ y $r = 10$. Los tiempos de absorción nos revelan la estructura global del grafo con las seis cuevas. Mientras menor sea el tiempo de absorción significa que el vértice correspondiente es más

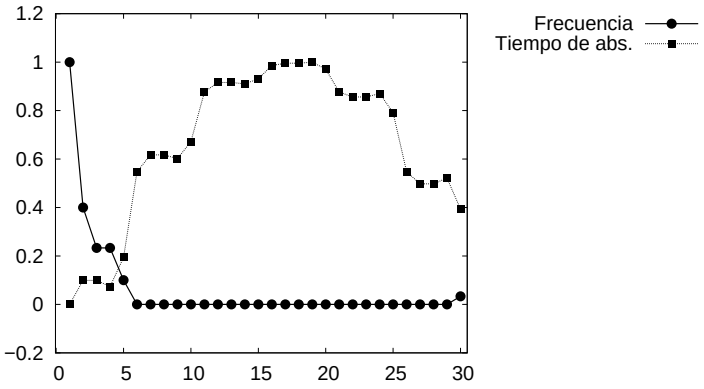


Figura 6.4: En el eje y están los tiempos de absorción y las frecuencias normalizadas mientras el eje x corresponde a los vértices.

cercano al vértice semilla, el cual siempre va a tener el mínimo tiempo de absorción, mientras que las frecuencias separan localmente el grupo del vértice semilla del resto del grafo, en este caso un mayor valor de frecuencia significa que el vértice es más cercano al vértice semilla, así la máxima frecuencia corresponde al vértice semilla. Tanto el vector de tiempos de absorción como el vector de las frecuencias han sido normalizados al intervalo $[0, 1]$.

Como medida de relación entre f_i y m_i se utilizó la *correlación* entre los tiempos de absorción y las frecuencias de visita; ésta se calculó con los valores originales de ambos vectores, es decir, no se normalizaron. La correlación es una medida de dependencia lineal que toma valores en el intervalo $[-1, 1]$. Cuando está en $(0, 1]$ se dice que existe una correlación positiva y ambos datos crecen linealmente, cuando está en $[-1, 0)$ la correlación entre los datos es negativa y mientras unos datos crecen los otros decrecen. Así mismo, cuando la correlación entre dos grupos de datos es 1 o -1 se dice que los datos están relacionados perfectamente y cuando es cero implica que no existe correlación alguna entre los datos. En nuestro caso de que exista algo de correlación ésta debe ser negativa pues recordemos que a tiempos de absorción pequeños le corresponden frecuencias altas. Para el grafo de la Figura 3.1, se muestra en la Figura 6.5 la correlación para cada uno de los seis grupos.

Otro ejemplo es un grafo generalizado de hombre de cueva [45] de 30 vértices y 248 aristas donde hay cuatro cuevas con densidad interna 0.95, es decir, 95 % de las posibles parejas ordenadas de vértices están conectadas, y la densidad entre las cuevas es 0.08. Las aristas fueron orientadas al azar para obtener un grafo dirigido. Se realizarón 30 repeticiones del método propuesto con $k = 2$ y $r = 1, 2, 3, \ldots, 30$. Los resultados obtenidos se muestran en la Figura 6.6. Observamos que el número de repeticiones para obtener una buena correlación es baja, con $r = 3$ se obtiene una buena correlación, lo cual da indicios de que el método que proponemos para agrupamiento local es eficiente.

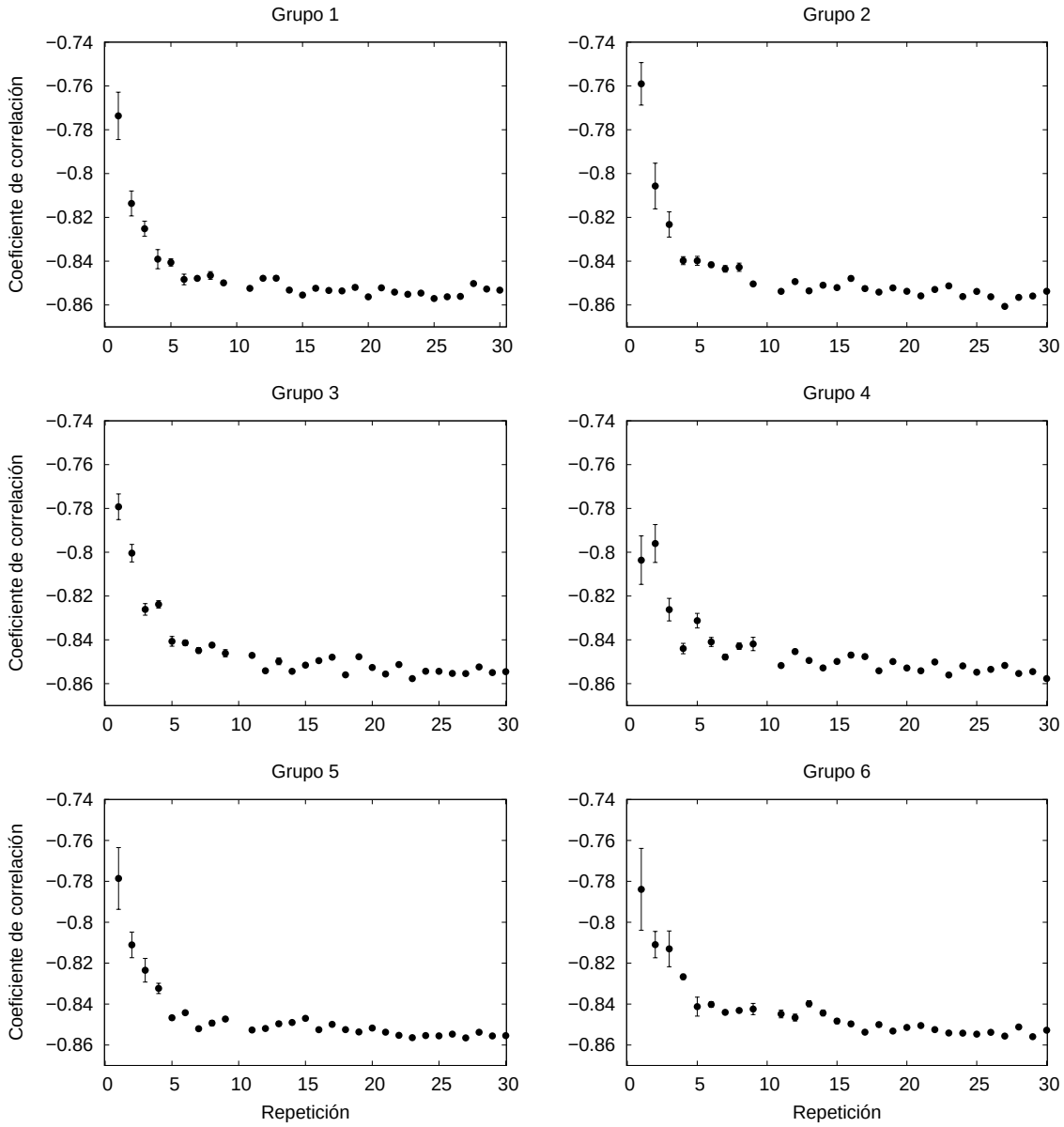


Figura 6.5: Correlación entre los tiempos de absorcíon y las frecuencias de visitas en el grafo de hombre de cueva (Figura 3.1) con parámetros $k = 2$ y $r = 1, 2, 3, \ldots, 30$, promedio y desviación estándar sobre 30 repeticiones del método propuesto de caminatas aleatorias cortas repetidas. Cada gráfica corresponde la correlación sobre el conjunto de vértices de cada una de las seis cuevas, utilizando cada uno como vértice semilla.

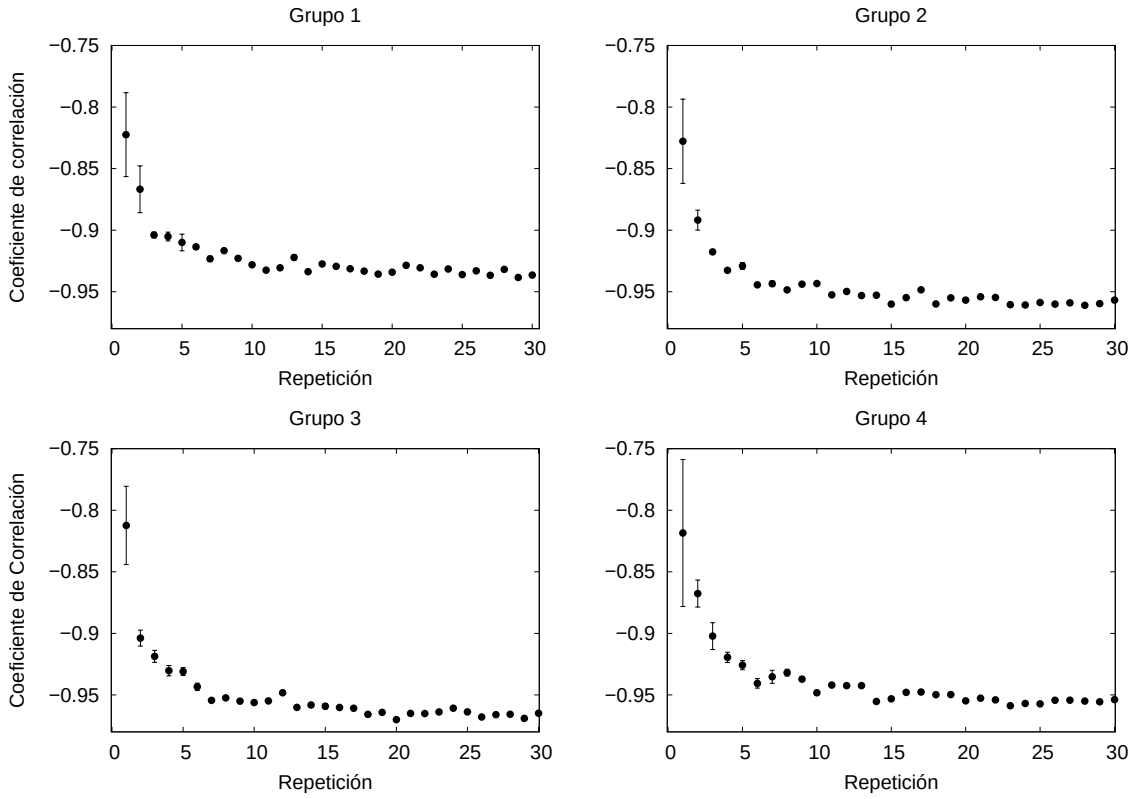


Figura 6.6: Correlación entre los tiempos de absorción y las frecuencias de visita en el grafo dirigido de hombre de cueva generalizado con $k = 2$ y $r = 1, 2, 3, \ldots, 30$, el promedio y desviación estándar son sobre las 30 repeticiones del método propuesto. Cada una de las gráficas muestra la correlación sobre el conjunto de vértices de cada una de las cuatro cuevas, utilizando cada uno como un vértice semilla.

Capítulo 7

Conclusiones

En el desarrollo de esté trabajo hemos presentado un método para el agrupamiento local de grafos dirigidos, el cual es motivado por la teoría y el agrupamiento espectral de grafos y su relación con los tiempos de absorción de caminatas aleatorias obtenidos al modelar los grafos dirigidos como una cadena de Markov. Durante todo el trabajo hemos hecho hincapié en que queremos encontrar localmente el grupo de vértices al cual pertenece un vértice semilla de interés, de tal modo que los vértices en tal grupo sean estructuralmente cercanos a la semilla. Del mismo modo también hemos visto que a un grafo dirigido se le puede asociar una cadena de Markov que corresponde a una caminata aleatoria ciega en el grafo.

Hemos aprovechado eficientemente el hecho de poder expresar la cercanía estructural en terminos de los tiempos de absorción para detectar vértices que son "relativamente cercanos" al vértice semilla; así mismo también hemos logrado detectar el grupo del mismo vértice semilla a través de caminatas aleatorias cortas repetidas desde este vértice, lo anterior se consiguió analizando la frecuencia de visitas a los otros vértices.

Los experimentos realizados se han hecho con grafos pequeños para tener la facilidad de comparar los resultados obtenidos con tiempos de absorción aproximados y la frecuencia de visitas de las caminatas aleatorias, con los tiempos de absorción exactos. Podemos concluir que el método desarrollado es eficiente, pues basta con una cantidad baja de repeticion de caminatas muy cortas para determinar al grupo

de vértices al cuál pertenece un vértice semilla dado.

Con lo anterior confirmamos la hipótesis supuesta, podemos hacer uso del vector de Fiedler para aproximar los tiempos de absorción. Esta forma de hacer agrupamiento local es una buena alternativa para el caso de grafos dirigidos debido a que los resultados existentes de agrupamiento global para el caso de grafos simples no aplican para ésta aplicación ya que la matriz Laplaciana asociada a un grafo dirigido no es simétrica por lo que dejan de cumplirse algunas propiedades importantes. En éste trabajo se propuso como solución alternativa el uso de los tiempos de absorción y las caminatas aleatorias.

Bibliografía

[1] Noga Alon. Eigenvalues and expanders. *Combinatorica*, 6(2):83–96, 1986.

[2] Reid Andersen, Fan R. K. Chung y Kevin Lang. Local partitioning for directed graphs using PageRank. En *Proceedings of WAW 2007*, páginas 166–178, 2007.

[3] Reid Andersen, Fan R. K. Chung y Kevin Lang. Local partitioning using PageRank vectors. En *Proceedings of the Fourty-seventh Annual Symposium on Foundations of Computer Science (FOCS)*, páginas 475–486, Washington, DC, EE.UU., 2006. IEEE Computer Society Press.

[4] Vanesa Avalos-Gaytán, Mario Rivera-Ramirez y Elisa Schaeffer. Agrupamiento local en grafos dirigidos. *Ciencia UANL, en revisión.*

[5] Pavel Berkhin. Survey of clustering data mining techniques. Informe técnico, Accrue Software, San Jose, CA, EE.UU., 2002.

[6] J.C. Bezdek, W.Q. Li, Y. Attikiouzel y M. Windham. A geometric approach to cluster validity for normal mixtures. *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 1(4):224–227, Diciembre 1997.

[7] J. A. Bondy y U. S. R. Murty. *Graph Theory with Applications.* Elsevier Science, 1976.

[8] Fan R. K. Chung. Laplacians and the Cheeger inequality for directed graphs. *Annals of Combinatorics*, 9(1):1–19, Abril 2005.

[9] Fan R. K. Chung. Four proofs of the Cheeger inequality and graph partition algorithms. En *Proceedings of the ICCM*, volumen II, páginas 1–4, 2007.

[10] Fan R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, Providence, RI, EE.UU., 1997.

[11] Fan R. K. Chung. Random walks and local cuts in graphs. *Linear Algebra and its Applications*, 423(1):22–32, Mayo 2007.

[12] Fan R.K. Chung y Robert B. Ellis. A chip-firing game and Dirichlet eigenvalues. *Discrete Mathematics*, 257:341–355, 2002.

[13] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest y Clifford Stein. *Introduction to Algorithms*. McGraw-Hill Book Co., Boston, MA, EE.UU., segunda edición, 2001.

[14] D.L. Davies y D.W. Bouldin. A cluster separation measure. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1(2):224–227, 1979.

[15] Reinhard Diestel. *Graph Theory*, volumen 173 de *Graduate Texts in Mathematics*. Springer-Verlag GmbH, Nueva York, NY, EE.UU., tercera edición, Julio 2005.

[16] Wilm E. Donath y A. J. Hoffman. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17(5):420–425, Septiembre 1973.

[17] J.C. Dunn. Well-separated clusters and optimal fuzzy partitions. *Cybernetics and Systems*, 4(1):95–104, 1974.

[18] Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23:298–305, 1973.

[19] Miroslav Fiedler. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Mathematical Journal*, 25:619–633, 1975.

[20] Vanesa Avalos Gaytán. *Segmentación de imágenes utilizando técnicas espectrales*. Tesis de licenciatura, Facultad de Ciencias Físico Matemáticas, Universidad Autónoma de C, Mayo 2007.

[21] E. N. Gilbert. Random graphs. *Annals of Mathematical Statistics*, 30(4):1141–1144, Diciembre 1959.

[22] John A. Hartigan y Manchek A. Wong. Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[23] Desmond J. Higham, Gabriela Kalna y Milla Kibble. Spectral clustering and its use in bioinformatics. *Journal of Computational and Applied Mathematics*, 204(1):25–37, Julio 2007.

[24] Michael Holzrichter y Suely Oliveira. A graph based method for generating the fiedler vector of irregular problems. En *Proceedings of the 11 IPPS/SPDP'99 Workshops Held in Conjunction with the 13th International Parallel Processing Symposium and 10th Symposium on Parallel and Distributed Processing*, volumen 1586 de *Lecture Notes In Computer Science*, páginas 978–985, Londres, G.B., 1999. Springer-Verlag.

[25] Anil K. Jain, M. Narasimha Murty y Patrick J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, Septiembre 1999.

[26] Ravi Kannan, Santosh Vempala y Adrian Vetta. On clusterings — good, bad and spectral. *Journal of the ACM*, 51(3):497–515, 2004.

[27] Jens Keuchel y Christoph Schnörr. Efficient graph cuts for unsupervised image segmentation using probabilistic sampling and SVD-based approximation. Informe técnico TR-2003-009, Institut für Informatik, Universidad de Mannheim, Alemania. Octubre 2003.

[28] Jon M. Kleinberg y Steve Lawrence. The structure of the web. *Science*, 294(5548):1849–1850, Noviembre 2001.

[29] Gregory F. Lawler. *Introduction to Stochastic Processes*. Chapman and Hall/CRC, Boca Raton, FL, EE.UU., segunda edición, 2006.

[30] Marina Meila y William Pentney. Clustering by weighted cuts in directed graphs. En *Proceedings of the Seventh SIAM International Conference on Data Mining*, Philadelphia, PA, EE.UU., 2007. Society for Industrial and Applied Mathematics.

[31] Marina Meila y Jianbo Shi. A random walks view of spectral segmentation. En *Proceedings of the Eighth International Conference on Artificial Intelligence and Statistics (AISTATS 2001)*, 2001.

[32] Mark E. J. Newman y Michelle Girvan. Mixing patterns and community structure in networks. En Romualdo Pastor-Satorras, Miguel Rubi y Albert Diaz-Guilera, editores, *Statistical Mechanics of Complex Networks*, volumen 625 de *Lecture Notes in Physics*, páginas 66–87, Berlin, Alemania, 2003. Springer-Verlag GmbH.

[33] Mark E.J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[34] Mark E.J. Newman y Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004.

[35] Pekka Orponen y Satu Elisa Schaeffer. Local clustering of large graphs by approximate Fiedler vectors. En Sotiris Nikoletseas, editor, *Proceedings of the Fourth International Workshop on Efficient and Experimental Algorithms (WEA'05)*, volumen 3505 de *Lecture Notes in Computer Science*, páginas 524–533, Berlin/Heidelberg, Alemania, 2005. Springer-Verlag GmbH.

[36] Pekka Orponen, Satu Elisa Schaeffer y Vanesa Avalos-Gaytán. Locally computable approximations for spectral clustering and absorption times of random walks. *En revisión.*

[37] Huaijun Qiu y Edwin R. Hancock. Graph matching and clustering using spectral partitions. *Pattern Recognition*, 39(1):22–34, Enero 2006.

[38] Satu Elisa Schaeffer. Stochastic local clustering for massive graphs. En T. B. Ho, D. Cheung y H. Liu, editores, *Proceedings of the Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-05)*, volumen 3518 de *Lecture Notes in Computer Science*, pages 354–360, Berlin/Heidelberg, Alemania, 2005. Springer-Verlag GmbH.

[39] Satu Elisa Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.

[40] Jianbo Shi y Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–901, Agosto 2000.

[41] Jiří Šíma y Satu Elisa Schaeffer. On the NP-completeness of some graph cluster measures. En Jiří Wiedermann, Gerard Tel, Jarslav Pokorný, Mária Bieliková y Július Štuller, editores, *Proceedings of the Thirty-second International Conference on Current Trends in Theory and Practice of Computer Science (Sofsem 06)*, volumen 3831 de *Lecture Notes in Computer Science*, páginas 530–537, Berlin/Heidelberg, Alemania, 2006. Springer-Verlag GmbH.

[42] Alistair J. Sinclair y Mark R. Jerrum. Approximative counting, uniform generation and rapidly mixing markov chains. *Information and Computation*, 82(1):93–133, Julio 1989.

[43] Daniel A. Spielman y Shang-Hua Teng. Spectral partitioning works: planar graphs and finite element meshes. En *Proceedings of the Thirty-seventh IEEE Symposium on Foundations of Computing (FOCS)*, páginas 96–105, Los Alamitos, CA, EE.UU., 1996. IEEE Computer Society Press.

[44] Benno Stein, Sven Meyer zu Eissen y Frank Wißbrock. On cluster validity and the information need of users. En *Proceedings of the Third IASTED In-*

ternational Conference on Artificial Intelligence and Applications, Septiembre 2003.

[45] Satu Elisa Virtanen. Properties of nonuniform random graph models. Research Report A77, Helsinki University of Technology, Laboratory for Theoretical Computer Science, Espoo, Finlandia, Mayo 2003.

[46] Duncan J. Watts. *Small Worlds*. Princeton University Press, Princeton, NJ, EE.UU., 1999.

[47] Fang Wu y Bernardo A. Huberman. Finding communities in linear time: a physics approach. *The European Physical Journal B*, 38(2):331–338, 2004.

[48] Wayne W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.

Índice de figuras

1.1. Diferentes objetos clasificados por propiedades en común. 1

1.2. Un agrupamiento de tres grupos, $d(A, C)$ y $d(B, C)$ denotan la distancia entre los grupos y $d(i, j)$ es la distancia entre objetos de un grupo. 2

2.1. Grafo simple. 8

2.2. El grado para cada uno de los vértices del grafo son: $d_1 = 2$, $d_2 = 2$, $d_3 = 3$, $d_4 = 1$. 9

2.3. En un grafo completo, cada vértice tiene grado $n - 1$. 9

2.4. En este ejemplo existe un camino entre cada par de aristas 10

2.5. Para un grafo dirigido como el que se muestra el grado de los vértices es: $d_1 = 0$, $d_2 = 2$, $d_3 = 2$, $d_4 = 1$. 10

2.6. Los vértices y aristas en negrita forman el subgrafo A. 11

2.7. Grafo simple conexo de seis nodos y su matriz de incidencia. 12

3.1. Un grafo con estructura de cuevas [46] compuesto de seis grupos los cuales están formados por cinco vértices cada uno, y que están conectados en un grafo circular el cual se forma al "remover" una arista de cada grupo para usarla como la arista que conecta al grupo con su grupo vecino. La arista que se remueve se muestra con una línea punteada. 16

3.2. Red social del club de karate [48]. Los dos grupos en que se dividió el club son indicados por la forma en que los vértices están dibujados: los cuadros representan un grupo y los círculos otro. 16

3.3. Componentes del vector de Fiedler (a la izquierda) y el vector de Fiedler normalizado (a la derecha) para el grafo de cuevas de la Figura 3.1. A simple vista la estructura de los seis grupos es evidente en el vector de Fiedler; en cambio, el vector de Fiedler normalizado agrupa los vértices en cuatro grupos, dos de ellos formados por dos cuevas. . 18

3.4. Los vértices estan graficados en orden correspondiente a los vértices en la Figura 3.2. Las componentes del vector de Fiedler (a la izquierda) y el vector de Fiedler normalizado (a la derecha) para la red del club de karate de la Figura 3.2. Los vértices pueden ser clasificados en dos grupos: aquellos cuyo valor es positivo en el vector de Fiedler y aquellos cuyo valor es negativo. 19

3.5. El objetivo del agrupamiento local es determinar el grupo al que pertenece un vértice de interés [35], por ejemplo en este grafo en amarillo denotamos el vértice de interés y los vértices con el borde más negro son los que pertenecen a su grupo. 20

4.1. La secuencia de vértices {1,2,3,4,5,6,7} es un camino. 23

4.2. La matriz de tiempos de absorción compuesta por 30 vectores de tiempos de absorción usando cada vértice del grafo de la Figura 3.1 como un vértice semilla. En blanco se representa al m_{ij} máximo, en negro al m_{ij} mínimo y ceros en la diagonal. 26

6.1. Comparación de los valores aproximados y exactos (ecuación 4.9) de los tres ejemplos de grafos: el de hombre de cueva de la Figura 3.1, el grafo del club de karate Figura 3.2 y el grafo $\mathcal{G}_{n,p}$, usando un vértice aleatorio como vértice semilla. En la primera fila mostramos el espectro ordenado de cada grafo, en la segunda la correlación entre el valor del vector exacto y el aproximado, y en la tercer y cuarta fila mostramos los valores del vector exacto y aproximado. 38

6.2. La suma de cuadrados exacta de la diferencia de los tiempos de absorción (a la izquierda) de los vectores estimados con diferentes valores para la aproximación a través de la ecuación 5.6 y la correlación de Pearson entre los valores de los vectores exactos y aproximados (a la derecha para los tres ejemplos de grafos: el grafo de las Figuras 3.1 y 3.2), y el grafo $\mathcal{G}_{n,p}$. Los valores mostrados son las medias sobre conjuntos de vértices de los dos primeros grafos, y sobre un conjunto de 100 vértices seleccionados uniformemente al azar para el grafo $\mathcal{G}_{n,p}$. La desviación estándar más pequeña corresponde al grafo de hombre de cuevas y la más grande al grafo aleatorio uniforme. Las líneas horizontales (las tres se traslapan entre 0.980 y 0.997) corresponden a las medias de los coeficientes de correlación entre los tiempos de absorción exactos y aproximados de la ecuación 5.7. 39

6.3. Cuatro ejemplos de dos-clasificación de vértices del la red del club de karate. Los ejemplos de la columna izquierda tienen vértice semilla representado por los rectángulos de la Figura 3.2 y los ejemplos de la columna derecha tiene vértice semilla representado por los círculos. Los vértices son ordenados en el mismo orden que muestra su etiqueta en la Figura 3.2 de la página 16 y en la posición correspondiente al vértice semilla se ha insertado un cero para representar absorción instantánea. El grupo al cual pertenece al vértice semilla está dibujado en color negro y el otro grupo en blanco. 40

6.4. En el eje y están los tiempos de absorción y las frecuencias normalizadas mientras el eje x corresponde a los vértices. 41

6.5. Correlación entre los tiempos de absorcíon y las frecuencias de visitas en el grafo de hombre de cueva (Figura 3.1) con parámetros $k = 2$ y $r = 1, 2, 3, \ldots, 30$, promedio y desviación estándar sobre 30 repeticiones del método propuesto de caminatas aleatorias cortas repetidas. Cada gráfica corresponde la correlación sobre el conjunto de vértices de cada una de las seis cuevas, utilizando cada uno como vértice semilla. 43

6.6. Correlación entre los tiempos de absorción y las frecuencias de visita en el grafo dirigido de hombre de cueva generalizado con $k = 2$ y $r = 1, 2, 3, \ldots, 30$, el promedio y desviación estándar son sobre las 30 repeticiones del método propuesto. Cada una de las gráficas muestra la correlación sobre el conjunto de vértices de cada una de las cuatro cuevas, utilizando cada uno como un vértice semilla. 44

Ficha autobiográfica

Vanesa Avalos Gaytán

Candidato para el grado de Maestra en Ciencias en Ingeniería
con especialidad en Ingeniería de Sistemas

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Tesis:

Agrupamiento Local en Grafos Dirigidos

Nací el 29 de abril de 1984. En el período 1999–2001 realicé mis estudios de preparatoria en la Escuela de Bachilleres: Doctor Mariano Narváez Gonzáles. En el año 2001 ingresé a la Facultad de Ciencias Físico Matemáticas, ambas de la Universidad Autónoma de Coahuila. En la Facultad realicé la tesis *Segmentación de imágenes utilizando técnicas espectrales* para obtener el título de Licenciada en Matemáticas Aplicadas en mayo del 2007. A partir de enero del 2007 soy estudiante del Programa de Posgrado en Ingeniería de Sistemas.