Minería de Redes

Alvaro J. Riascos Villegas

Octubre de 2022

Contenido

Redes Sociales

- Una red social es un grafo pero típicamente con una característica notoria de no aleatoridad o localidad.
- Intuitivamnete si A y B y A y C están relacionados probablemente B y C también lo están.
- Otra característica importante de estas redes es la conformación de comunidades (i.e., un individuo puede pertenecer a varias comunidades). Lo que es distinto a particiones.
- Estudiaremoes tres formas de detección de comunidades: métodos estándar de agrupamiento, *betweenness* y el método basado en el Laplaciano.

Redes Sociales como Grafos

- Intuitivamente es esta red un ejemplo de red social?

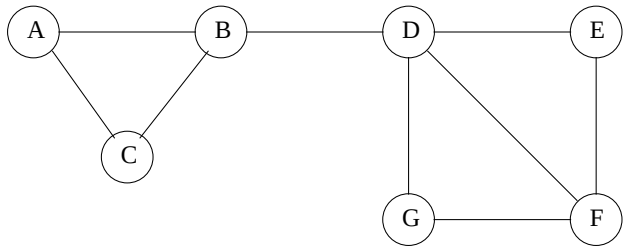


Figure 10.1: Example of a small social network

- Suponga que X, Y, Z son nodos en la red, si X, Y y X, Z están enlazados, cuál es la probabilidad de que Y y Z también lo esten?
- En un grafo **cualquiera** con 7 nodos existen $\binom{7}{2} = 21$ enlaces posibles.
- Ahora, en un grafo **cualquiera con** 7 **nodos y** 9 **enlaces** (no necesariamente el de la figura), si sabemos que los nodos X, Y, y X, Z arbitrarios estan enlazados entonces la probabilidad de que Y y Z también lo esten es: $\frac{9-2}{21-2} = 0{,}368$.

Redes Sociales como Grafos

- Intuitivamente es esta red un ejemplo de red social?

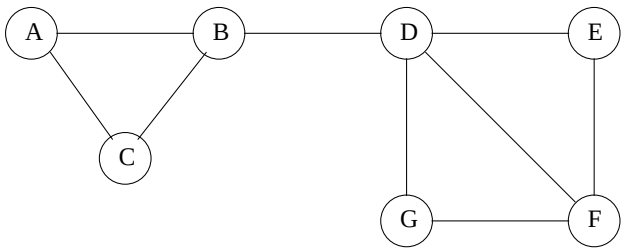


Figure 10.1: Example of a small social network

- Ahora calculemos esta probabilidad específicamente para la red propuesta. El resultado es $\frac{9}{16} = 0{,}563$ que es bastante mayor que el cáculo anterior.
- Concluimos que esta red si tiene las características de una red social.

Ejemplo: Empleabilidad

- Granovetter (1960) entrevistó a muchas personas que recientemente habian cambiado de trabajo. La mayoría de las personas lo hizo mediante contactos personales, pero no directos sino *conocidos*.
- Los sociologos racionalizan esto suponiendo que los contactos cercanos tienen la misma información y solo los contactos menos directos traen nueva información (oportunidades, etc.).
- Granovetter formaliza esto mostrando que la mayoría de los contactos que resultaron en trabajos son *puentes locales débiles*.

Puentes

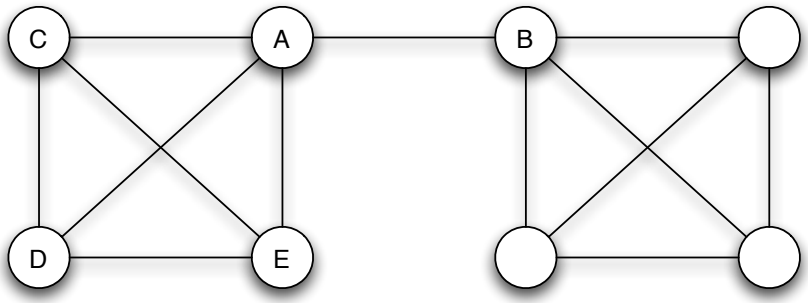


Figure 3.3: The A-B edge is a *bridge*, meaning that its removal would place A and B in distinct connected components. Bridges provide nodes with access to parts of the network that are unreachable by other means.

Puentes locales

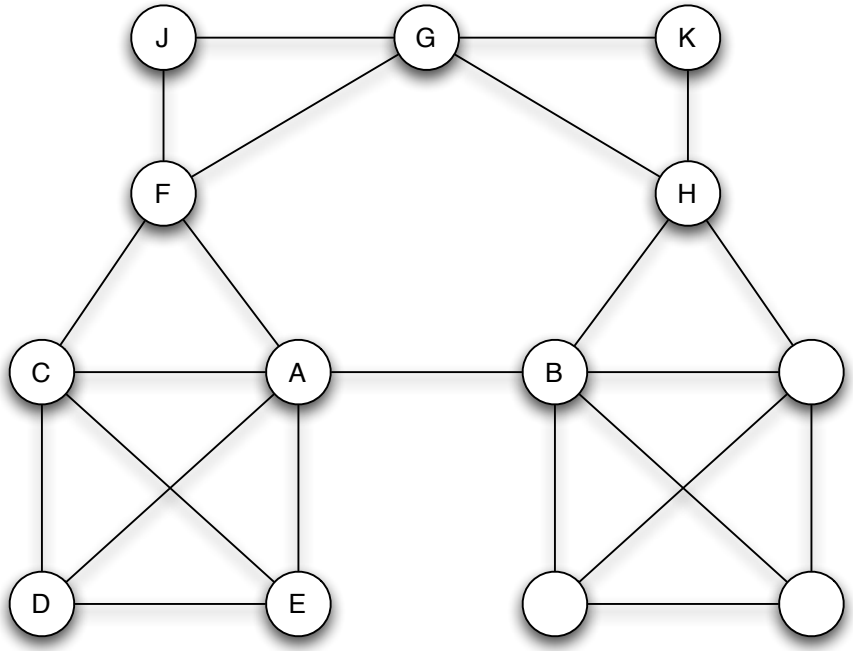


Figure 3.4: The A-B edge is a local bridge of span 4, since the removal of this edge would increase the distance between A and B to 4.

- Un enlace es un puente local si une dos nodos que no tienen ningun vecino en común (eliminarlo aumenta en tres o más la distancia entre los nodos).

Enlaces fuertes y débiles

- Supongamos que es posible marcar todos los enlaces como fuertes o débiles.

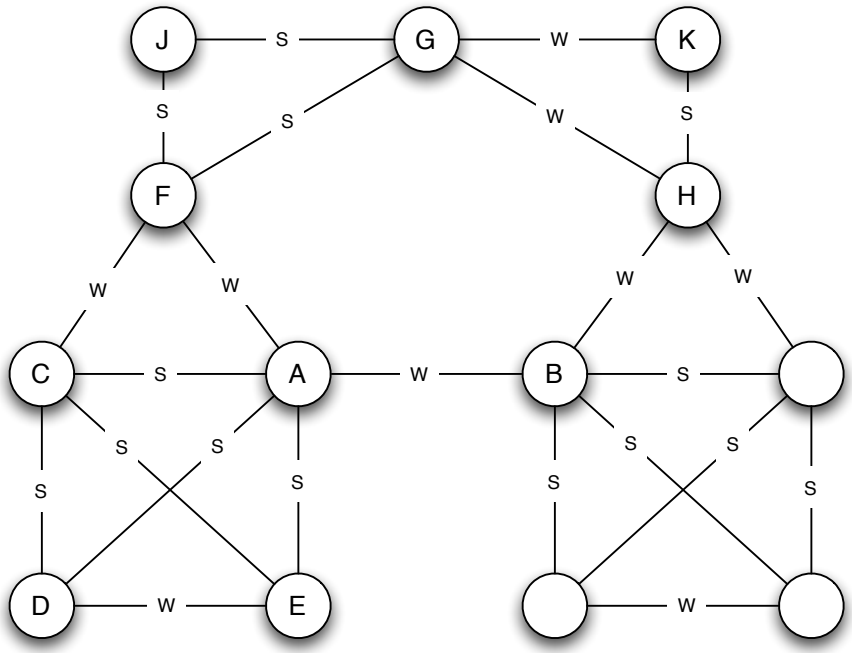


Figure 3.5: Each edge of the social network from Figure 3.4 is labeled here as either a *strong tie* (S) or a *weak tie* (W), to indicate the strength of the relationship. The labeling in the figure satisfies the Strong Triadic Closure Property at each node: if the node has strong ties to two neighbors, then these neighbors must have at least a weak tie between them.

Relación entre puentes locales y enlaces débiles

- Los puentes reflejan características globales un grafo. El concepto de enlaces fuertes o débiles características locales.
- La clave para relacionar estos dos características de un grafo es el concepto de clausura triádica.

Clausura triádica

- Clausura triádica. Cuando dos personas tiene un amigo en común pero ellas mismas no lo son, existe una posibilidad alta de que se vuelvan amigas.
- El coeficiente de aglomeración local es una forma de cuantificar el concepto de clausura triádica.
- Bearman y Moody [2004]. Suicide and friendship among American adolescents. Encuentran que mujeres adolescentes con menos coeficiente de aglomeración local son más propensas a suicidarse.

Clausura triádica: fuerte

- Decimos que A viola la propiedad de clausura triadica fuerte si tiene enlaces fuertes con B y C y no existe ningun enace entre B y C (de ningún tipo). Decimos que A satisface la propiedad si no la viola.
- En la siguiente figura todos los nodos satisfacen la propiedad.

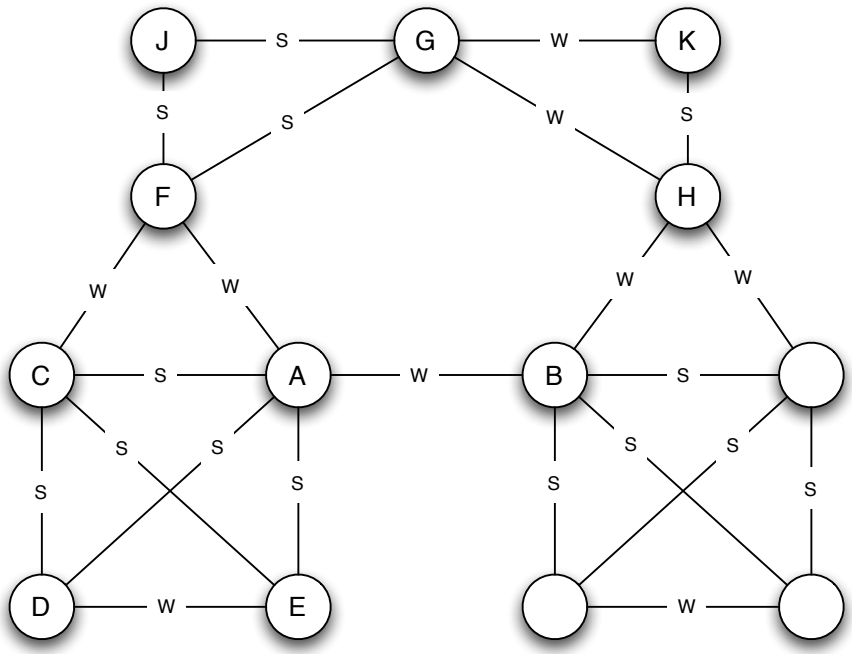
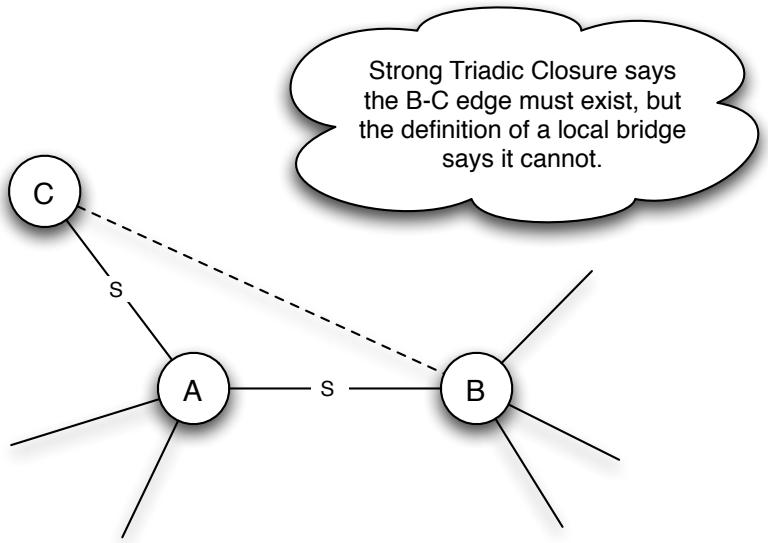


Figure 3.5: Each edge of the social network from Figure 3.4 is labeled here as either a *strong tie* (S) or a *weak tie* (W), to indicate the strength of the relationship. The labeling in the figure satisfies the Strong Triadic Closure Property at each node: if the node has strong ties to two neighbors, then these neighbors must have at least a weak tie between them.

Clausura triádica: fuerte

- **Afirmación**: Si un nodo en una red satisface la propiedad de clausura triádica fuerte y tiene por lo menos dos vecinos con enlaces fuertes entonces en cualquier puente local que este involucrado, este debe ser un enlace débil.



Puntes locales y enlaces débiles

- El resultado anterior captura la idea de que en redes sociales, donde probablemente existen muchos nodos que satisfacen la propiedad de clausura triádica fuerte, los puente locales son enlaces débiles (conocidos).
- Estos conocidos juegan un papel importante acercando personas con enlaces débiles.

Contenido

Métodos de segmentación estándar

- El agrupamiento jerárquico va tener dificultades cuando considere los nodos By D.
- Una dificultad similar tendría el agrupamiento basado en técnicas de asignación de puntos (e.g., K-medias).

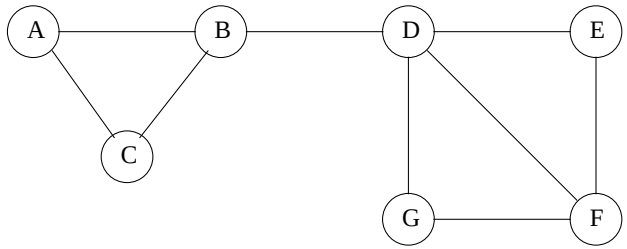
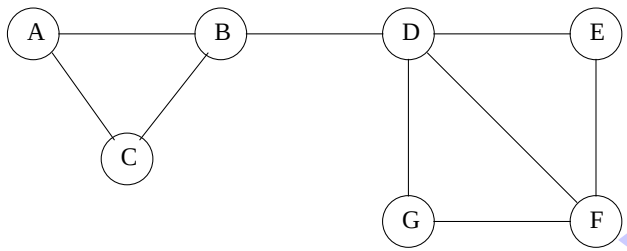


Figure 10.1: Example of a small social network

Contenido

Betweenness

- Definimos la intromisión de un enlace x, como el número de parejas de nodos $\{a, b\}$ tal que el enlace x se intromete entre los caminos más cortos entre a y b. Como pueden existir varios caminos más cortos entre a y b se le asigna a x solo el la fracción que pasan por x.
- Entre mayor sea este puntaje, más indicativo de que el enlace separa dos grupos.
- En la figura abajo (B, D) tiene un puntaje de $3 \times 4 = 12$, y (D, F) tiene un puntaje de 4.
- Calcular este puntaje puede ser difícil en redes grandes.



Betweenness: Algoritmo de Girvan-Newman

- Paso (1): Representación BFS del ejemplo anterior comenzando en E.

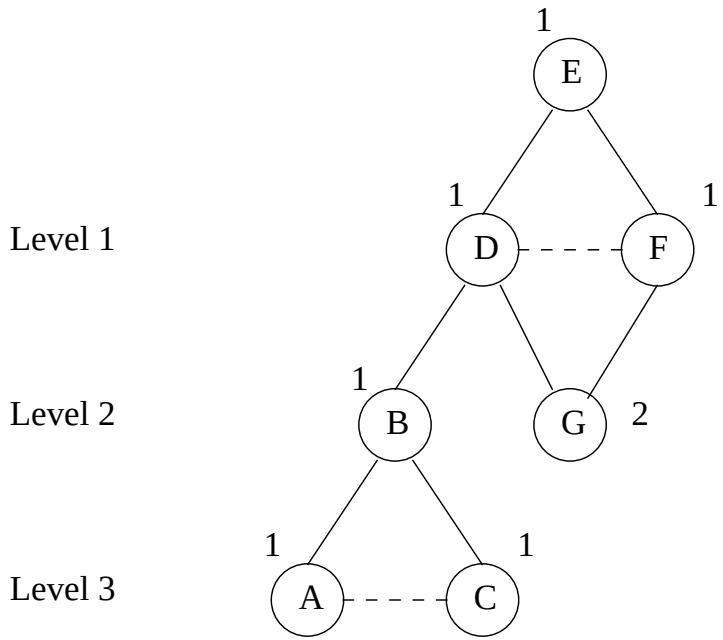


Figure 10.4: Step 1 of the Girvan-Newman Algorithm

- Las líneas continuas se llaman DAG y las punteadas representan nodos en el mismo nivel y no pueden hacer parte de caminos más cortos comenzando en E.

Betweenness: Algoritmo de Girvan-Newman

- Paso (2): Marcar cada nodo con la cantidad de caminos más cortos que llegan al nodo comenzando en E.

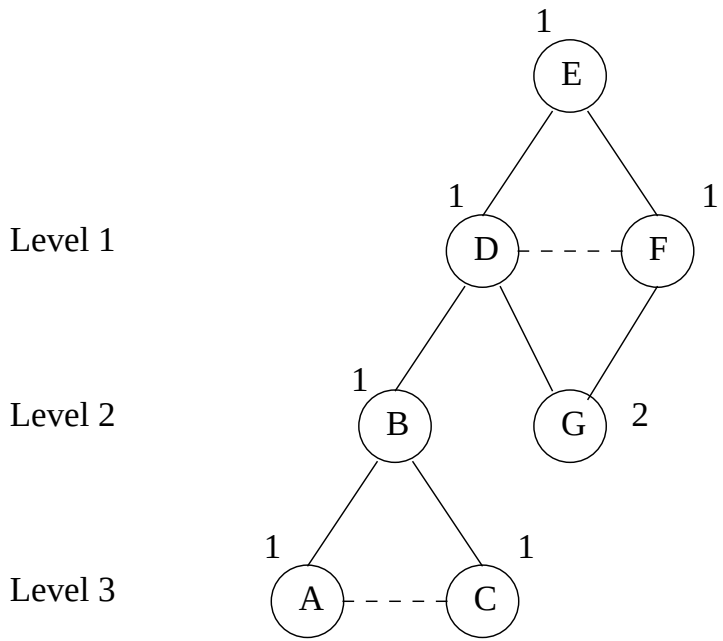


Figure 10.4: Step 1 of the Girvan-Newman Algorithm

Betweenness: Algoritmo de Girvan-Newman

- Paso (3): Cada enlace va recibir un peso de acuerdo a la siguientes reglas:
  - Cada hoja del grafo se le da un crédito de 1 (modificando lo que tuviera en el paso anterior si es necesario).
  - Cada nodo que no sea una hoja recibe 1 más el peso de los enlaces que le suceden.
  - Los pesos de los enlaces se heredan del peso de los nodos que los suceden (de forma proporcional).

Betweenness: Algoritmo de Girvan-Newman

- Paso (3a): Marcar enlaces BA y BC y nueva marca del nodo B y G resultado del primer paso de inducción.

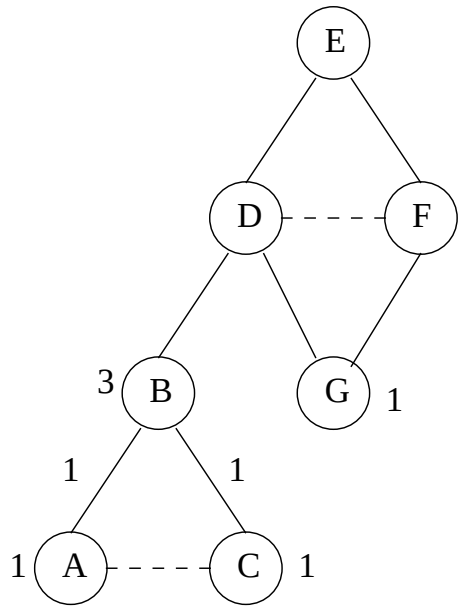


Figure 10.5: Final step of the Girvan-Newman Algorithm – levels 3 and 2

Betweenness: Algoritmo de Girvan-Newman

- Paso (3b): Marcar siguientes enlaces compartiendo el puntaje proporcional a el peso de los nodos parientes.

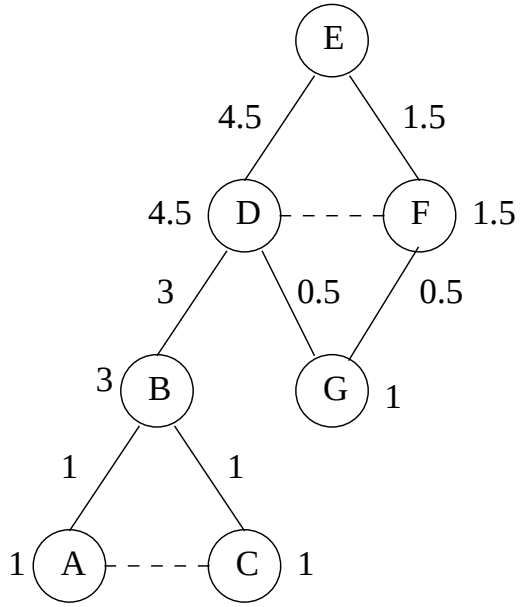


Figure 10.6: Final step of the Girvan-Newman Algorithm – completing the credit calculation

- El resultado final es un puntaje para cada enlace dada la raíz y la representación BFS.
- Obsérvese que que cada enlace queda con el puntaje adecuado de intromisión de todos los caminos más cortos que comienzan en X.

Betweenness: Algoritmo de Girvan-Newman

- Paso (4): Repetir este procedimiento para cada nodo como raíz y su representación BFS.
- Sumar el puntaje de cada enlace que se obtiene en cada representación.
- Dividir por dos dado que cada camino se va descubrir dos veces, una vez para cada nodo inicial y terminal de un camino más corto.

Betweenness: Algoritmo de Girvan-Newman

- Resultado final de aplicar el algoritmo al ejemplo.

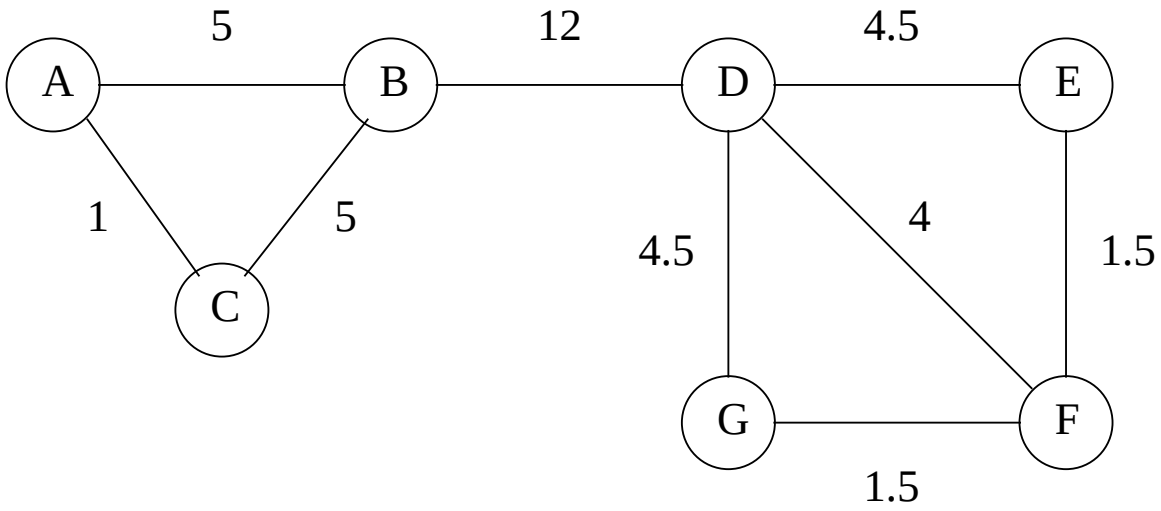


Figure 10.7: Betweenness scores for the graph of Fig. 10.1

Betweenness y detección de comunidades

- Calcular el puntaje de intromisión de cada enlace y comenzar a eliminar elaces desde el puntaje más alto hasta obtener una segmentación adecuada.

Betweenness y detección de comunidades

- Comunidades a partir de la eliminación de todos los puntajes superiores a o iguales a 4.

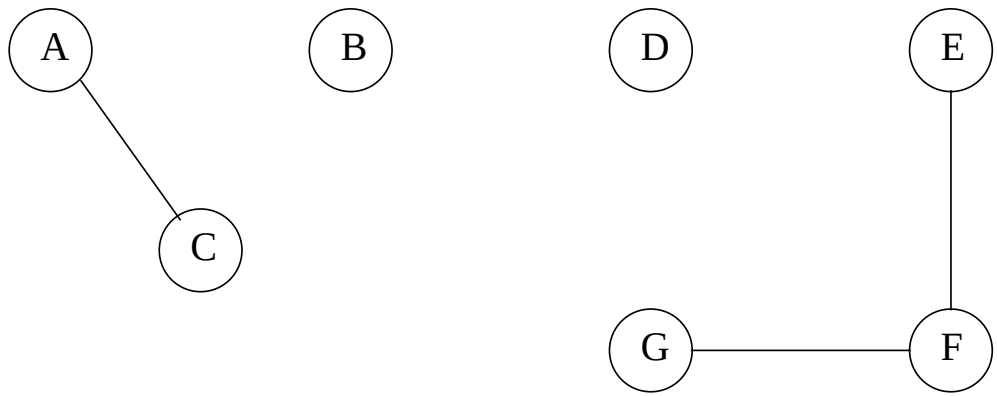


Figure 10.8: All the edges with betweenness 4 or more have been removed

Contenido

Comunidades

- Los algoritmos descritos hasta el momento particionan los grafos. No permiten que un nodo pertenezca a varias comunidades.
- La idea del método a continuación es identificar nodos que tiene mucho enlaces.
- Una primera idea es buscar *cliques*. Sin embargo, encontrar cliques grandes es un problema muy difícil computacionalmente (i.e., NP-Completo).
- Resulta que en los grafos bipartitos grandes siempre se puede encotrar un subgrafo bipartito completo.

Comunidades

- Estos puden ser utilizado como el punto de partida (i.e., núcleo) de la construcción de una comunidad: dado el núcleo añadir nodos que tiene muchos enlaces con el núcleo.
- Si el grafo no es bipartito se puede dividir aleatoriamente los nodos en dos clases.
- A continuación estudiamos como encontrar subgrafos bipartitos completos.

Subgrafos bipartitos completos

- Denotamos por $K_{s,t}$ un grafo completo de s nodos de una clase y t nodos de otra.
- Sea G un grafo bipartito. El objetivo es econtrar instancias $K_{s,t}$ grandes en G.
- Esto se puede ver como un problema de encontrar items frecuentes en canastas de consumo (i.e., análisis de canastas).
- Tomemos una de las clases de G como el conjunto de items (s) y la otra clase como el conjunto de canastas (t). Una canasta la describe un nodo y los items asociados. Supongamos que hay menos canastas que items ($t \leq s$).

Subgrafos bipartitos completos

- Visto como un problema de encotrar canastas la idea es: buscar t items que estén en s canastas. Esto es una instancia de $K_{s,t}$.
- En el siguiente ejemplo hay dos instancias de $K_{2,1}$.

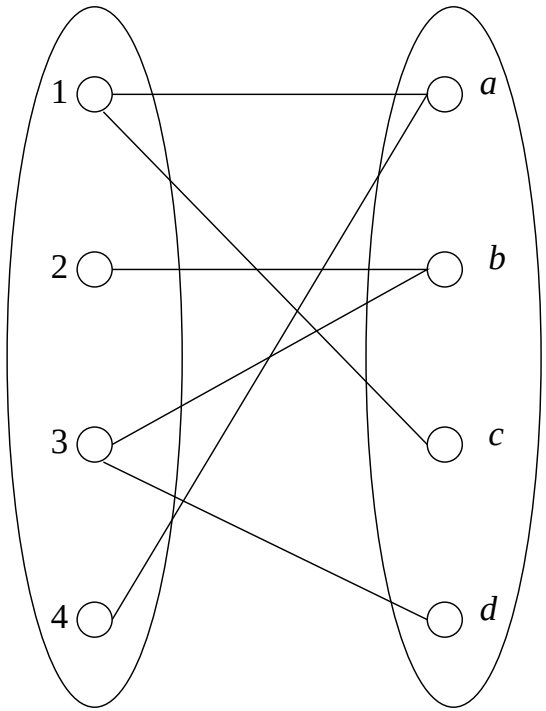


Figure 10.10: The bipartite graph from Fig. 8.1

- Se puede demostrar que siempre se pueden encontrar instancias de $K_{s,t}$.

Contenido

Particiones

- La idea es particionar los nodos de tal forma que se minimize los enlaces que conectan estas componentes pero teneiendo en cuenta el tamaño de cada componente.
- Definimos el volúmen de un conjunto de nodos S como el número de enlaces con al menos un (end) nodo en S.
- Sea S, T dos conjuntos de nodos disyuntos y $Cut(S, T)$ el número de enlaces de S a T.
- Definimos el corte normalizado como:

$$Cut(S, T)(\frac{1}{Vol(S)} + \frac{1}{Vol(T)})$$

Particiones

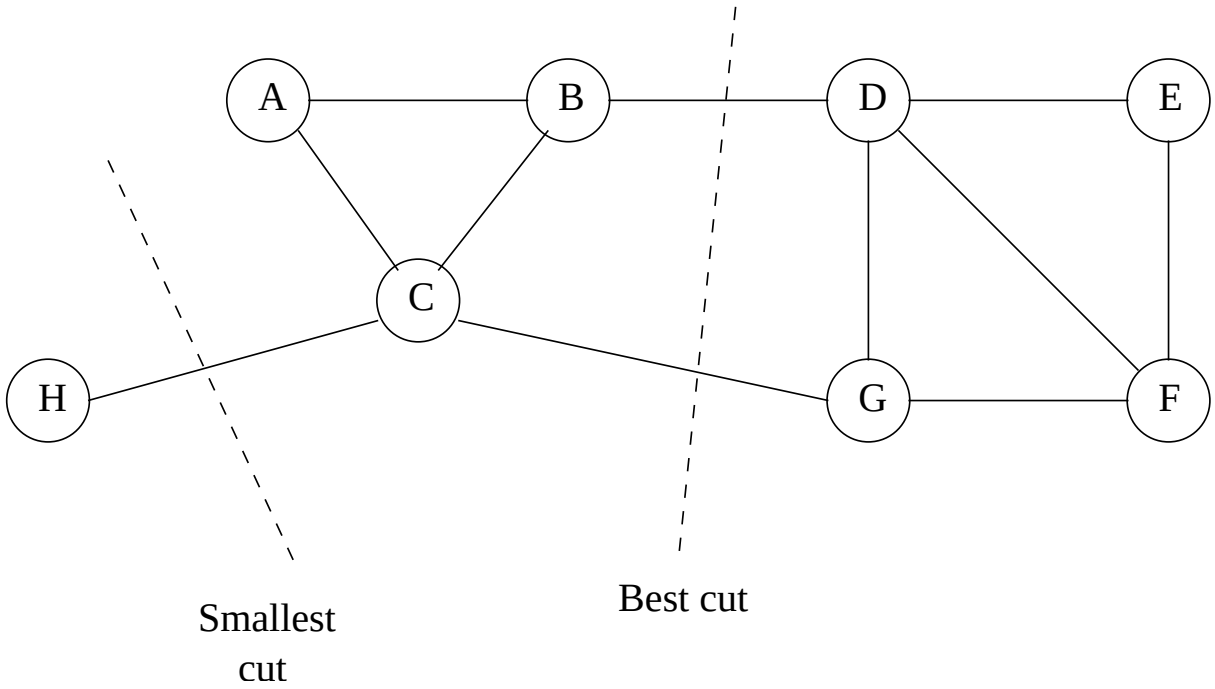


Figure 10.11: The smallest cut might not be the best cut

- En esta red, el corte normalizado de la izquierda es: $\frac{1}{1} + \frac{1}{11} = 1{,}09$ y el de la derecha es: $\frac{2}{6} + \frac{2}{7} = 0{,}62$.

Laplaciano

- Definimos el Laplaciano L de una red como: $L = D - A$, donde A es la matriz de adjacencia (con ceros en la diagonal) y D es la matriz diagonal con el grado de cada nodo en la diagonal.
- Ejemplo:

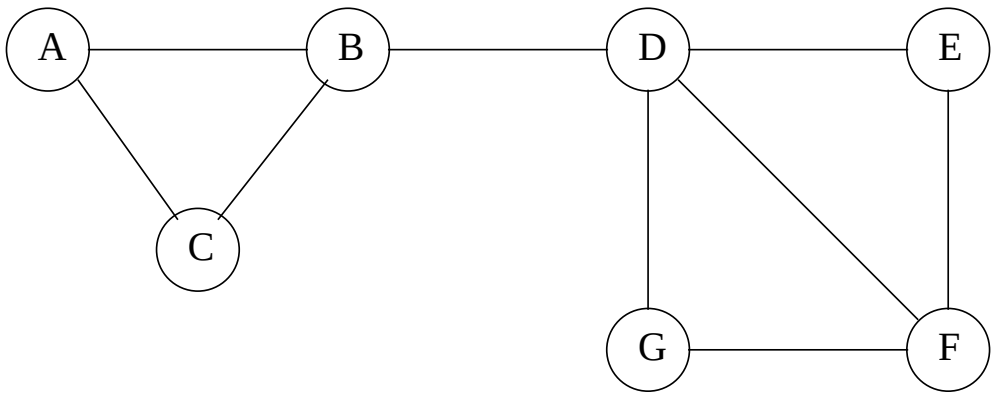


Figure 10.1: Example of a small social network

Laplaciano: Valores propios

- Los valores y vectores propios del Laplaciano contienen información global de un grafo.
- El menor valor propio del Laplaciano es cero (con vector propio el vector de unos).
- Ahora obsérvese que $x^T L x = \sum_{\{i,j\}:A_{ij}=1}(x_i - x_j)^2$ y consideremos el problema:

$$\min_{\|x\|_2=1, x \perp \mathbf{1}} x^T L x^T \tag{1}$$

Se puede demostrar que la solución a este problema es igual al segundo menor valor propio de L y, el x que resuelve el problema, el vector propio asociado.

Laplaciano: Valores propios

- Es fácil de ver que no todas las componentes de x pueden ser cero, algunas deben ser positivas y otras negativas y la tendencia es a que tengan el mismo signo si existe un enlace entre los nodos.
- Esto motiva el siguiente método para particionar un grafo.

Ejemplo: Laplaciano

- Ejemplo:

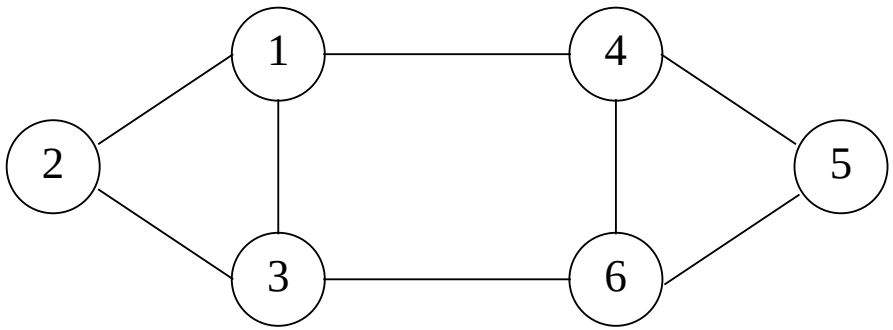


Figure 10.16: Graph for illustrating partitioning by spectral analysis

Ejemplo: Laplaciano

- Ejemplo:

$$\begin{bmatrix} 3 & -1 & -1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & 0 & 0 & -1 \\ -1 & 0 & 0 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & -1 & -1 & 3 \end{bmatrix}$$

Figure 10.17: The Laplacian matrix for Fig. 10.16

Ejemplo: Laplaciano

- Ejemplo:

Eigenvalue	0	1	3	3	4	5
Eigenvector	1	1	−5	−1	−1	−1
	1	2	4	−2	1	0
	1	1	1	3	−1	1
	1	−1	−5	−1	1	1
	1	−2	4	−2	−1	0
	1	−1	1	3	1	−1

Figure 10.18: Eigenvalues and eigenvectors for the matrix of Fig. 10.17

Contenido

Caminata aleatoria con reinicio

- El objetivo es estimar qué tan similares son los nodos dentro de una misma categoría a un nodo específico.
- No es posible medir la similaridad entre dos nodos simplemente navegando de forma aleatoria en un grafo pues la distribucion final es independiente del punto de partida (estado estacionario de PageRank).
- Dado un nodo de interes, la idea consiste en usar el algoritmo de Pagerank pero con una probabiliad positiva de teletransportarse concentrada en el nodo de interes.
- El resultado arroja una medida de similitud de cada nodo al nodo de interes. Esto calculo hay que hacerlo para cada nodo.

Ejemplo

Web Page Title	John McCarthy's View PageRank Percentile	Netscape's View PageRank Percentile
John McCarthy's Home Page	100.00%	99.23%
John Mitchell (Stanford CS Theory Group)	100.00%	93.89%
Venture Law (Local Startup Law Firm)	99.94%	99.82%
Stanford CS Home Page	100.00%	99.83%
University of Michigan AI Lab	99.95%	99.94%
University of Toronto CS Department	99.99%	99.09%
Stanford CS Theory Group	99.99%	99.05%
Leadershape Institute	95.96%	97.10%

Table 2: Page Ranks for Two Different Views: Netscape vs. John McCarthy