Análisis de redes sociales

Gonzalo Rivero

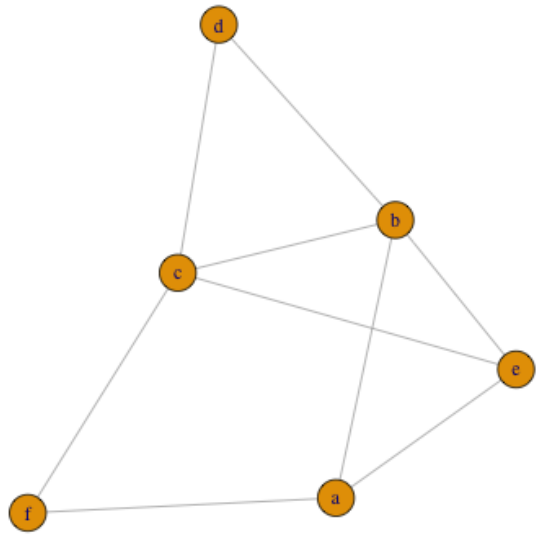
Big Data en investigación social y opinión pública

22 a 26 de julio, 2019

¿Qué son las redes sociales?

- Una forma de pensar sobre los sistemas sociales que pone la atención sobre las relaciones entre las entidades que forman el sistema.
- Actores (nodos) tienen características (atributos) y están unidos por relaciones (aristas) que pueden tener características propias.
- Las relaciones crean cadenas o caminos indirectos entre diferentes nodos.
- Los actores no tienen por qué estar limitados a agentes sociales.

Grafo de una red social



Los nodos (p.e., a) puede representar personas (Antonio) con atributos (47 años) y las aristas representar relación (de amistad) con atributos propios (durante 4 años). Algunas relaciones son indirectas (a no conoce a d si no a través de b)

Objetivos del análisis de redes

- Estudios descriptivos:
  - Centralidad de drogadictos para intervención social con objeto de que las prácticas se difundan.
  - Estudiar comunidades dentro de una fusión de empresas que no se integran en la nueva organización.
- Estudios multivariados:
  - Similitudes en actitudes o comportamiento que da lugar relaciones de amistad.
  - Tipo de cultura de organización que predice creación de confianza.

Niveles de análisis

- Análisis a nivel de diada, nodo y red.
  - *Diada:* Probabilidad de formar amistad depende de cercanía de sus despachos en la empresa.
  - *Nodo:* Centralidad en la red de confianza en la empresa predice ascensos.
  - *Red:* Los caminos más cortos en la red de comunicación del grupo predice la capacidad de resolver problemas.

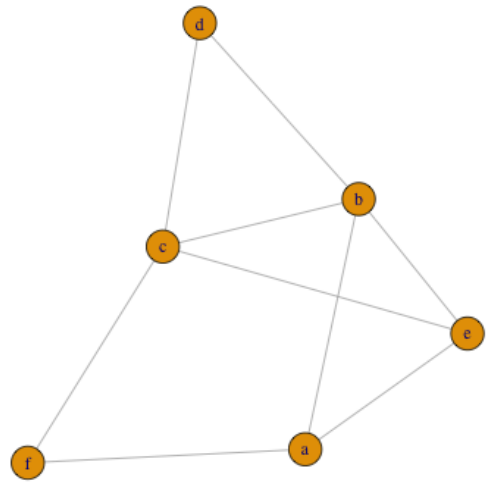
Tipos de análisis

- Medidas de la red pueden ser variable dependiente o independiente.
- Podemos usar la red para entender resultados.
  - Agentes económicos bien conectados políticamente tienen mejores resultados.
- También podemos estudiar cómo atributos de la red pueden depender de factores exteriores
  - Entender *homofilia* (predisposición a relacionarse con agentes similares) en medios sociales.

Redes como grafos

- Forma más habitual de conceptualizar la red es como un grafo $G(V, E)$ que contiene vértices (nodos o puntos) V y aristas (enlaces o relaciones) E.
- Las aristas conectan pares de vértices adyacentes.
- Grafos pueden ser dirigidos si aristas son pares ordenados.
- Podemos tener más de una relación en un grafo (multiplex). Por ejemplo, relaciones de amistad y vínculo familiar.

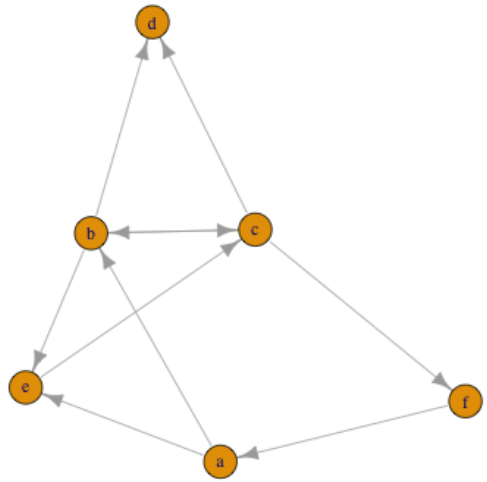
Grafos no dirigidos



En esta red

$$V = \{a, b, c, d, e, f\}$$
$$E = \{\{a, b\}, \{b, c\}, \{c, d\}, \dots\}$$

Grafos dirigidos



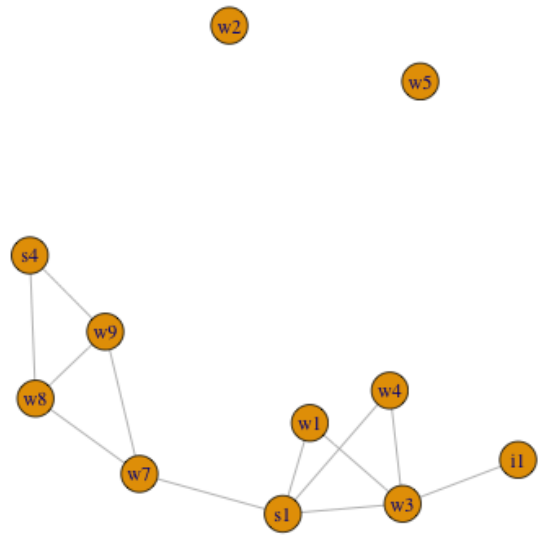
En esta red

$$V = \{a, b, c, d, e, f\}$$
$$E = \{(a, b), (b, c), (c, b), (c, d), \dots\}$$

Caminos

- Un *path* es una secuencia de nodos adyacentes.
- Si el grafo es dirigido, la secuencia debe respetar la dirección de las aristas.
- Para ser un *path*, la secuencia no puede repetir vértices.
- Si repite nodos pero no repite aristas es un *trail*.
- Cualquier secuencia de nodos adyacentes es un *walk*.
- La longitud de un *walk* es el número de aristas que recorre.
- La *path* más corto entre dos vértices es la *geodésica*.

Caminos

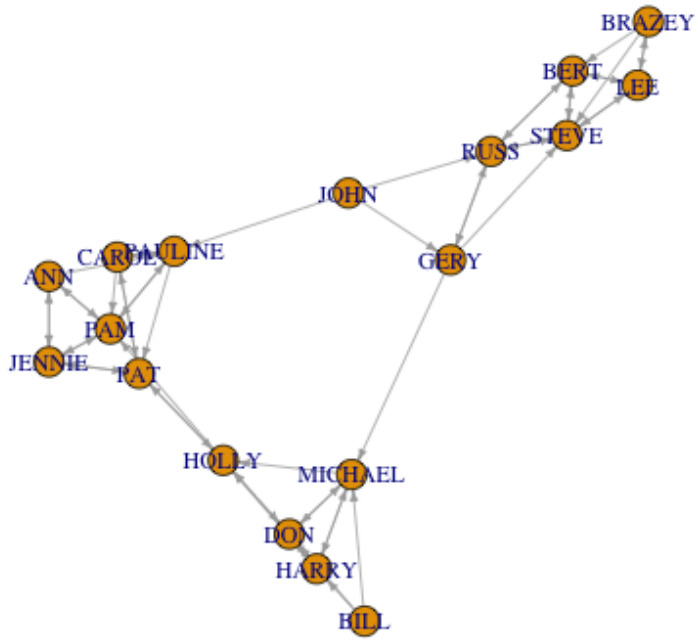


- $s4 - w9 - w8 - w7$ es un path.
- $w9 - w8 - s4 - w9 - w7$ no es un path (pero es un walk).
- $w9 - w8 - s4 - w9 - w7$ es un trail.
- $w8 - w7 - w9 - w7 - s1$ no es un trail (pero es un walk).
- $w3 - w1 - s1$ y $w3 - w4 - s1$ son geodésicas.

Componentes

- Algunos nodos no se pueden alcanzar por ningún *path*.
- Un *componente* es el máximo grupo de nodos que puede alcanzarse a través de un *path*. *Máximo* indica que no pueden añadirse nodos al componente sin violar la condición.
- La cercanía entre dos nodos es su "cohesión", que puede ser una arista o el peso de una arista.

Componentes



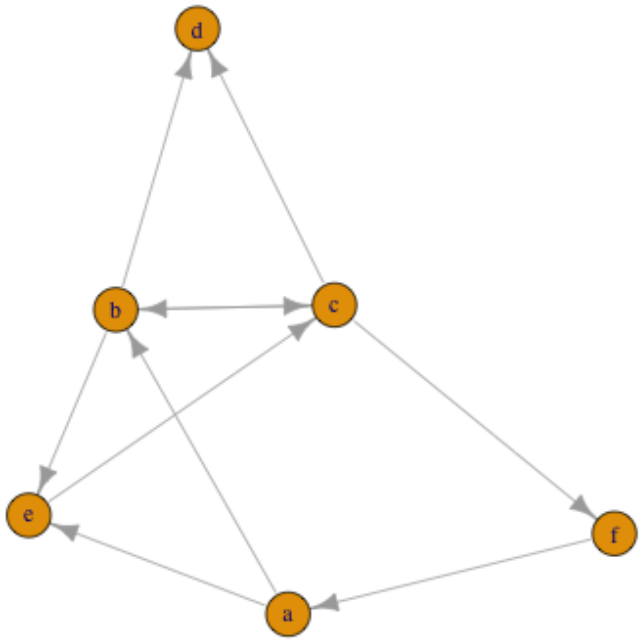
Lee, Bert y Brazey no son un componente, porque podríamos añadir Steve y todo el mundo podría llegar a otro. Pero Lee, Bert, Brazey, Steve, Russ y Gery son un componente.

Matrices de adyacencias

Podemos representar una red como una matriz.

- Filas y columnas representan nodos y las entradas representan aristas.
- Convención de que dirección va de filas a columnas. Eso nos permite representar grafos direccionales.
- Pueden existir bucles si la diagonal no es cero.

Matrices de adyacencias



	a	b	c	d	e	f
a	0	1	0	0	1	0
b	0	0	1	1	1	0
c	0	1	0	1	0	1
d	0	0	0	0	0	0
e	0	0	1	0	0	0
f	1	0	0	0	0	0

Ambos objetos representan la misma red. Nótese que d no tiene ningún enlace saliente, pero dos enlaces entrantes (de b y e).

Matrices de adyacencias

- Si F es matriz de amistades y E es relación de enemistades, FE es "enemigo de un amigo de" y cada elemento (i, j) es el número de amigos de i que tiene j como enemigo.
- Multiplicación de matrices no es simétrica: amigos de mis enemigos no son los enemigos de mis amigos.
- Multiplicación F^k es el número de *walks* de longitud k que empiezan en una fila y acaban en una columna: cuántos amigos tiene i que son amigos de j.

Diseño de estudios de redes

- La mayoría de los estudios tradicionales eran observacionales, mediante encuesta y con un único punto en el tiempo.
- Algunos experimentos y algunos estudios longitudinales.
  - Sujetos aleatorizados en diferentes condiciones con diferentes enlaces para estudiar cambios en cooperación.
  - Estudios longitudinales sobre contagio de la obesidad en el que se siguen a actores a lo largo del tiempo.

Diseños del estudio de redes

- Podemos diseñar estudios para recuperar toda la red (todas las aristas entre un conjunto de nodos) o redes personales (estudios de relación entre *ego* y sus *alters*).
- Estudios de red completa son más costosos pero dan una visión general sobre la estructura.
- Estudios de red personal respetan la confidencialidad pero no permiten entender la estructura completa. Además es más fácil definir los límites de la red al no tener que seguir a todos los nodos.

Consideraciones de diseño

- Muestreo en redes presenta dificultades:
  - Aclarar el tamaño de la red en la pregunta de investigación. Pueden usarse criterios internos o externos a la red.
    - Estudio de amistades en un club de karate
    - Influencia en decisiones de consumo
  - Alto impacto de los valores perdidos en nodos y aristas.
- Alto coste de las encuestas de red:
  - Que el entrevistado enumere los nodos con los que tiene relación
  - Entrevistas a los nodos nombrados por el entrevistado

Ética en estudio de redes

- Las redes presentan consideraciones éticas y de privacidad que no aparecen en otros estudios de encuestas.
  - Diseños de red completa no pueden ofrecer anonimato
  - Rechazo a participar como entrevistado no evita inclusión en la red.
  - Visualización puede deanonimizar información. Es posible recuperar la identidad de alguien a través de con quién está relacionado.

Visualización

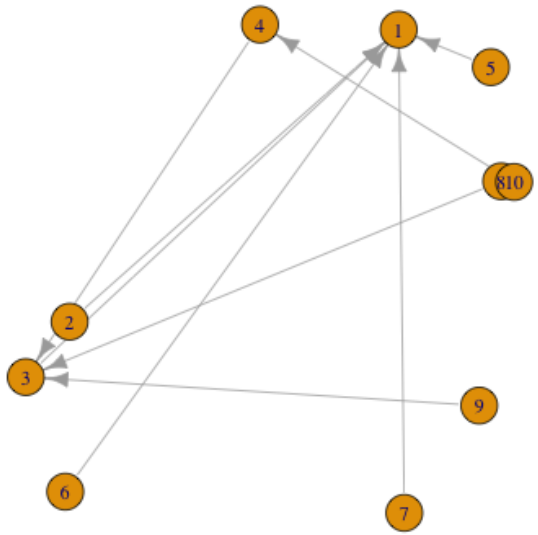


Figura: Uniforme en el plano

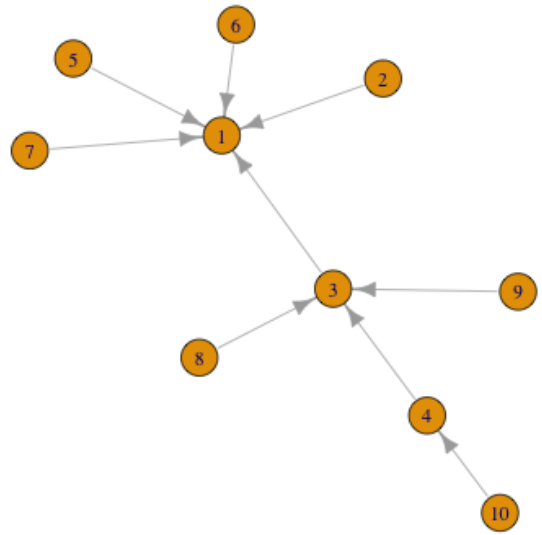


Figura: Fruchterman-Reingold

La disposición de los nodos puede ayudarnos con la interpretación. Ambos grafos representan la misma red.

Visualización

- El *diseño* de un diagrama se refiere a la posición de los vértices en el plano.
- Existen varias estrategias:
  - Aleatoriamente no suele ser buena idea.
  - Distribuirlos mediante atributos como en un gráfico de dispersión.
  - Con métodos de ordenación como MDS que reflejen la distancia entre diferentes nodos (lazos fuertes $\Rightarrow$ cercanía).
  - Algoritmos que optimicen criterios, como Fruchterman-Reingold
    - Correspondencia entre distancias.
    - Maximizar la distancia entre nodos.
    - Preferencia por aristas de igual longitud.

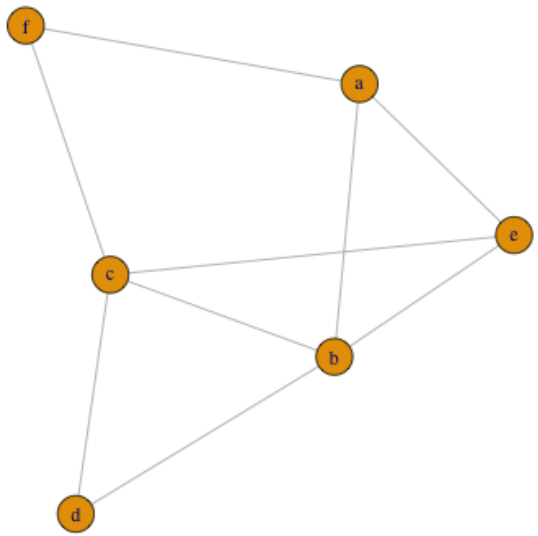
Descriptivos de una red

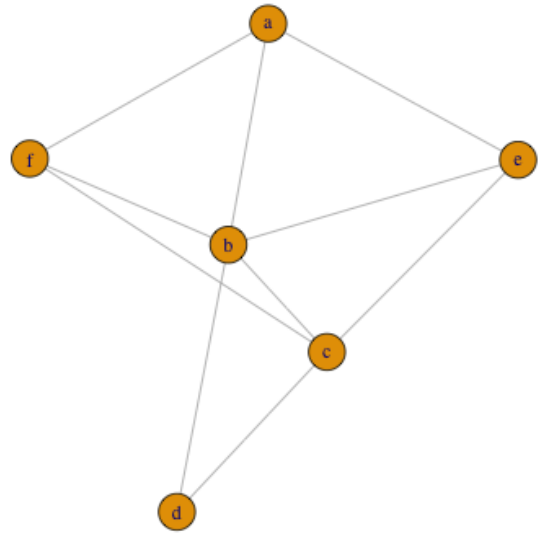
- Diferentes formas de caracterizar una la estructura red para describir su grado de *centralización*.
- Interpretación de qué significa *centralidad* es relativo a la pregunta que nos interesa.
- Dos objetivos diferentes: grado de centralización de la red y centralidad de cada nodo.

Medidas de cohesión

- Cohesión se refiere a grado en el que los nodos están conectados en la red o en un subgrupo.
- *Grado medio de la red*: Media del número de aristas adyacentes (entrada o salida) a cada nodo.
- *Densidad*: Número de aristas en la red sobre todas las posibles (probabilidad de que exista una arista entre dos nodos).
- *Conectividad*: Proporción de pares de nodos que pueden alcanzarse mutuamente a través de cualquier path. Es una medida de robustez a eliminación de nodos. Si poderamos por inversa a longitud es *compactación*.

Medidas de cohesión





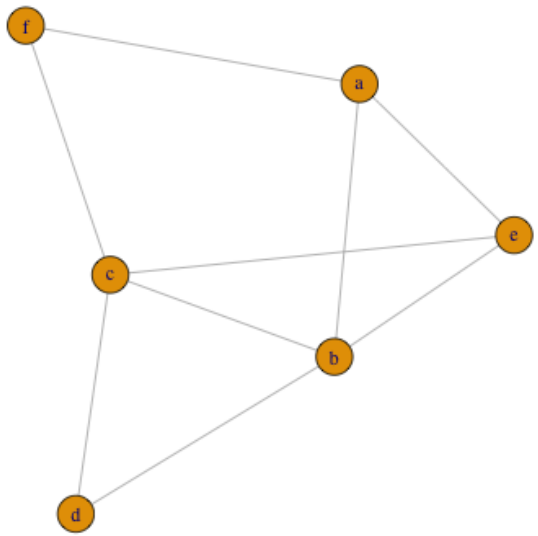
Densidad:	0.6
Grado medio:	3
Conectividad:	2

Densidad:	0.66
Grado medio:	3.3
Conectividad:	2

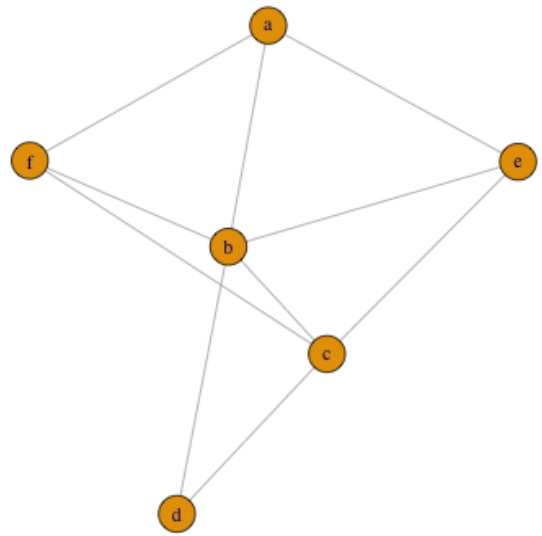
Medidas de transitividad

- Forma convencional de cuantificar cercanía: si A está relacionado con B y B está relacionado con C, ¿está A relacionado con C?
- Si hay muchos triángulos en la red, tiene estructura compacta.
- *Transitividad*: Triadas existentes sobre todas las posibles.
- *Coeficiente de conglomeración*: Media de la densidad de relaciones en la red de cada "ego" ponderada por número de pares de nodos. Idéntico a transitividad.
- Relación con "redes de mundos pequeños".
  - Obsevación de Watts y Strogaz de que solo 6 aristas son necesarias para conectarnos.
  - La red es muy compacta pero también es posible viajar rápido.
  - Solo es necesario tener aristas aleatorias entre pares de individuos.

Medidas de transitividad



Transitividad: 0.45



Transitividad: 0.57

Medidas de centralidad

- Objetivo es medir la posición de un nodo en la red.
- Diferentes medidas que capturan la contribución individual de cada nodo a la estructura de la red.
- Aproxima definiciones de "importancia" puede referirse a impacto si el nodo es eliminado o número de conexiones.
- Captura la ventaja de la que disfruta un nodo en virtud de su posición. En contextos sociales, ofrece información de recibir información y ejercer influencia.

Medidas de centralidad: centralidad en grado

- Número de enlaces que llegan al nodo y por tanto una medida local que no necesita de conocer el resto de la red. No puramente centralidad.
- Es una medida del grado de exposición: es probabilidad de alcanzar un nodo asumiendo un walk aleatorio.
  - Muchos procesos interesantes (flujo de información) no son walks aleatorios.
- En matriz de adyacencias: $d_i = \sum_j x_{ij}$

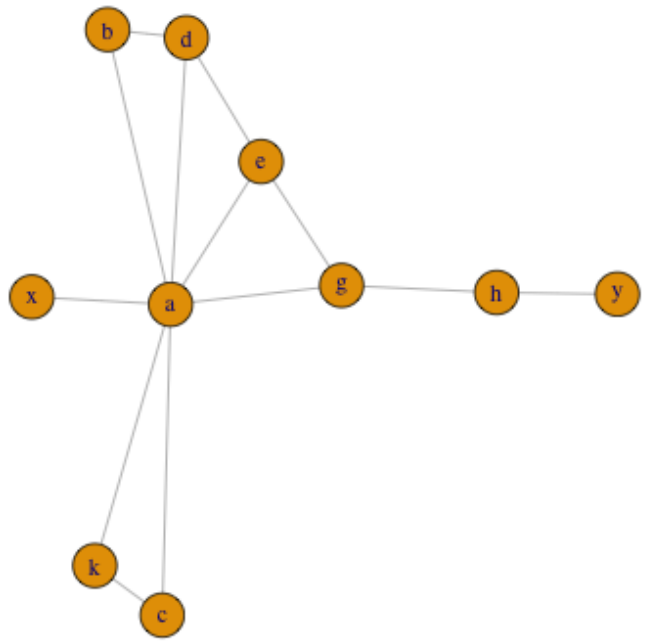
Medidas de centralidad: centralidad en autovector

- Similar a centralidad en grado pero ponderando cada nodo por su propia centralidad en grado.
- En matriz de adyacencias:

$$e_i = \lambda \sum_j e_j x_{ij} \tag{1}$$

  en donde λ es mayor autovector de la matriz de adyacencias.
- Es una formulación recursiva: un nodo es central si está conectado a nodos que son centrales.
- Forma simple del algoritmo PageRank de Google.

Centralidad en autovector



Los nodos x e y tienen mismo grado pero x está conectado a un nodo más popular. Mejor estimación de riesgo (por ejemplo de STD).

Medidas de centralidad: centralidad beta

- Generalización de centralidad en autovector y en grado,

$$c = \alpha(A1 + \beta A^2 1 + \beta^2 A^3 1 + \dots) \tag{2}$$

- Interpretación de las potencias k de la matriz de adyacencias (walks entre i y j de longitud k)
- β es escogido por el usuario. $\beta = 0$ es centralidad en grado. Converge a centralidad en autovector $\beta = 1/\lambda$
- β controla longitud de caminos en los que estamos interesados. Mayor β mayor longitud de los caminos que consideramos.
- Medida de influencia directa e indirecta, con influencia indirecta ponderada inversamente por distancia.

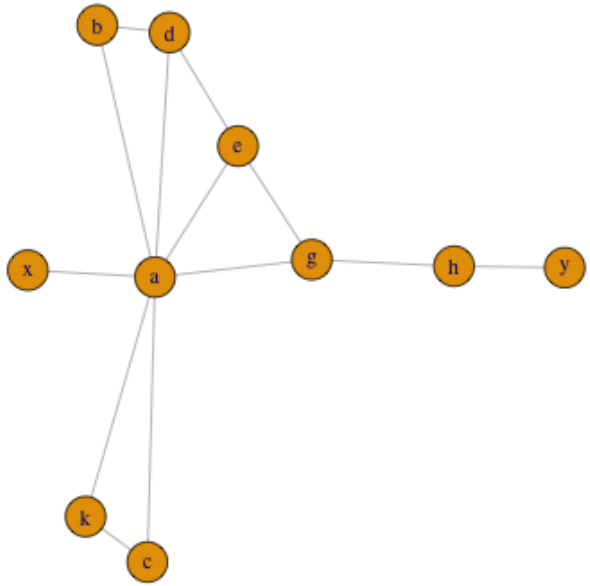
Medidas de centralidad: centralidad en intermediación

- Con qué frecuencia un nodo está en el camino más corto entre otros dos
- $b_j = \sum_{i<k} \frac{g_{ijk}}{g_{ik}}$ con g siendo el número de caminos geodésicos que conectan i con k (a través de j).
- Toma valor cero si nunca pertence al camino de menor distancia
- Interpretado como control de los flujos dentro de la red: nodos de alta intermediación pueden actuar de filtro o control.

Medidas de centralidad: centralidad en cercanía

- Suma de las geodésicas a todos los otros nodos. Habitualmente dividida por $n-1$ (valor mínimo).
- Captura el minimo tiempo de llegada de un flujo a través de la red. Un nodo con alta cercanía está a poca distancia de los demás y la información originada en él tarda poco en difundirse.

Comparación de medidas



	a	b	d	c	e	g	h	y	k	x
Btn	25.0	0.0	0.5	0.0	1.5	14.0	8.0	0.0	0.0	0.0
Eig	1	0.4	0.6	0.4	0.6	0.5	0.1	0.1	0.4	0.3
Cls	8.3	5.2	5.5	5.2	6.2	6.6	4.7	3.4	5.2	5.0

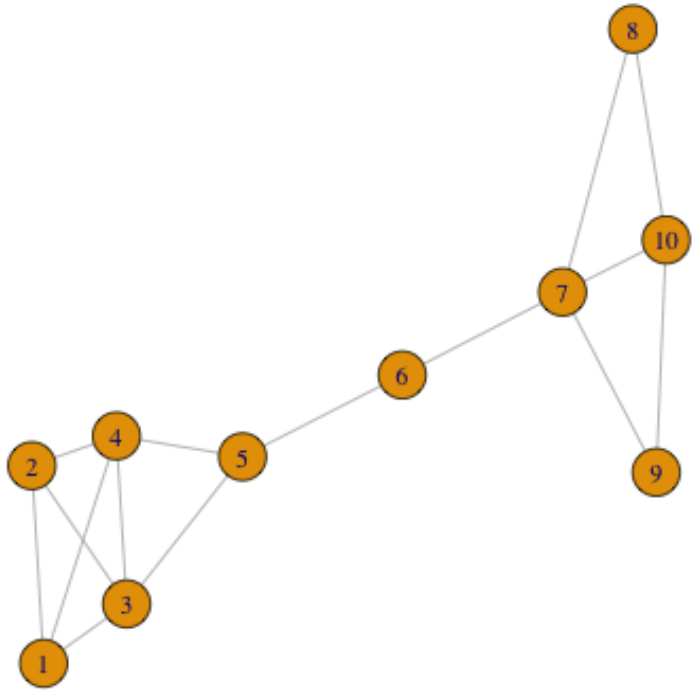
Comunidades

- Podemos pensar en subunidades dentro de la red que están relacionadas entre ellas pero no con los demás
- Primer objetivo es identificar la existencia de las comunidades
- Segundo objetivo es explicar qué características determinan estas comunidades

Cliques

- Un clique es un subconjunto de actores en los que cada actor es adyacente a todos los otros actores y es imposible añadir más actores sin violar esta condición
- Pertenencia a cliques comunes es medida de cercanía entre dos actores
- Varios métodos para identificar cliques superpuestos como forma de comunidad
- Pueden ser difíciles de identificar cuando hay muchos subgrupos sobrepuestos

Cliques



$\{1, 2, 3, 4\}$ son un clique, pero $\{1, 2, 3\}$ no

Identificación de comunidades

- Existen muchas alternativas dependiendo de la definición de comunidad. Una muy popular e intuitiva es el método de Girvan-Newman.
- Identificar aristas que si son eliminadas fragmentan la red con mayor intermediación
- Podemos seguir eliminando aristas hasta llegar a un número k predeterminado.

Identificación de comunidades

- Un método muy popular es la maximización de la modularidad. Detecta divisiones del grafo que tienen alta modularidad
- La modularidad compara los grupos con redes aleatorias: alta densidad de conexiones por encima de un grafo aleatorio.
- El método de Lovaina genera comunidades locales hasta que la modularidad global no puede ser mejorada.
- Métodos de más interes utilizan modelos generativos que indican cómo la red se ha formado y permite la realización de inferencia estadística.

Contraste de hipótesis

- Es posible hacer contraste de hipótesis en redes
  - Red de matrimonios en Florencia de Padgett y Ansell
  - Hipótesis es que familias que hacen negocios juntas también tienden a casarse
  - Calcular correlación entre dos matrices de adyacencias
- Problemas de aproximaciones habituales
  - Diseñadas para trabajar con vectores, no con matrices
  - Supuestos de independencia que no se cumplen
  - Distribuciones alejadas de la normal
- Dos soluciones
  - Usar métodos de permutación
  - Hacer modelo generativo de redes

Contrastes de permutación

- Lógica de permutación es calcular modos en los que el experimento podría haber ocurrido haciendo asignación aleatoria
- Técnica QAP (quadratic assignment procedure):
  - "Aplastar" matrices de adyacencias en dos vectores y calcular correlación observada
  - Reorganizar filas y columnas equivalentes y calcular correlaciones entre las nuevas matrices
  - Comprobar proporción de casos en los que la distribución anterior produce correlación observada.
- Es posible extenderlo a modelos de regresión

Modelos de grafos exponenciales aleatorios

- Definir un modelo base que define expectativa sobre los datos, similar a imponer un modelo lineal
- Supuesto de independencia condicional:
  - Si dos vértices comparten una arista son dependientes condicional al resto del grafo.
- Modelo origina un número pequeño de configuraciones esperados y las posibles configuraciones se escogen en función de una hipótesis.
- Comparar las configuraciones existentes con las esperadas.
  - Un parámetro de transitividad alto indica más triángulos de los que cabría esperar dada la densidad de la red y la propensión a formar triángulos.
- Puede estimarse mediante MLE