FACULTAD DE INGENIERÍA

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Métricas y Medidas

Paola Bermolen

paola@fing.edu.uy

1. **Medidas de centralidad**
   - Centralidad de grado
   - Centralidad de vector propio
   - Centralidad de cercanía
   - Centralidad betweenness
2. **Densidad local, clustering coefficient y centralidad de grupos**
3. **Assortativity mixing**
   - Modularidad

Importancia de los vértices

- En análisis de redes, muchas veces aparece la idea de importancia de un vértice

Ejemplo

- En una red social, ¿quién tiene “las riendas del poder”?
- ¿Cuánta autoridad tiene una página de la WWW considerada por sus “peers”?
- ¿Qué neuronas son más importantes para ciertas actividades del cerebro?
- ¿Qué importancia tiene un cruce de semáforos en los desplazamientos diarios?

Centralidad de grado

- Medidas de centralidad de vértices intentan cuantificar la noción de importancia
  - $\Rightarrow$ en redes chicas, se puede “ver” pero no así para redes grandes
  - $\Rightarrow$ permite comparar
  - $\Rightarrow$ fuertemente inspirado en *social network analysis*
- Se define el grado como medida de centralidad de un vértice
- Medida más simple pero ilustrativa
  - $\Rightarrow$ *influencers* en redes sociales
  - $\Rightarrow$ redes de citas

Centralidad por vector propio

- Idea: Un vértice es importante si sus vecinos son también importantes

  $\Rightarrow$ en la centralidad por grado, todos los vecinos son igualmente importantes

- Dado G un grafo simple, no dirigido, la centralidad por vector propio de un vértice i queda definida de manera implícita como

$$x_i = k^{-1} \sum_{(i,j) \in E} x_j \quad k \text{ constante de proporcionalidad}$$

- Una centralidad alta se consigue con pocos vecinos importantes o muchos vecinos no tan importantes

Centralidad por vector propio

- Usando la definición de matriz de adyacencia;

$$x_i = k^{-1} \sum_{(i,j)\in E} x_j = \sum_{i,j} A_{ij} x_j$$

- Si $X = (x_1, x_2, \cdots, x_n)^t$, en forma matricial resulta:

$$\mathbf{A}X = kX$$

- X es vector propio de A asociado al valor propio K.
- Imponiendo que $x_i \geq 0$ para todo $i \in V$, resulta que:
  - $\Rightarrow$ K es el valor propio más grande (dominante) de **A** [Bonacich'87]
- Si G es conexo, el teorema de Perron-Frobenius asegura que:
  - $\Rightarrow$ el valor propio más grande de **A** es positivo y de multiplicidad 1
  - $\Rightarrow$ todas las entradas del vector propio asociado X son positivas

Centralidad por vector propio

- Se puede calcular X y $K = \lambda_1$ con complejidad $O(N_v^2)$ método de las potencias

$$x_i(k+1) = \frac{\mathbf{A}x_i(k)}{\|\mathbf{A}x_i(k)\|},\ k = 0, 1, \ldots$$

- ¿Qué pasa con grafos dirigidos? sumamos importancia de los nodos con arcos incidentes

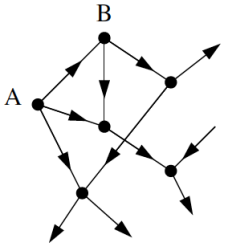
$$x_i = k^{-1} \sum_j A_{ji} x_j$$

$$A^t X = kX \quad \text{equivalente a } X^t A = kX^t$$

- X es vep a izquierda de A o vep de A^t asociado al valor propio dominante

Centralidad por vector propio en digrafos

- Se puede definir, pero no funciona bien...



- $x_A = 0$ (ningún nodo incidente)
- implica $x_B = 0$
- implica $x_C = 0$

- Centralidad no nula solo para nodos en una componente fuertemente conectada o su componente out

Centralidad Katz

- Katz (1953): cada nodo tiene un poco de importancia β de "regalo"

$$x_i = \alpha \sum_j A_{ji} x_j + \beta = \alpha \sum_j A^t_{ij} x_j + \beta$$

- Matricialmente: si $\mathbf{1} = (1, 1, \cdots, 1)^t$,

$$X = (I - \alpha A^t)^{-1} \beta \mathbf{1}$$

  $\Rightarrow$ ¿cómo elegir α? Se elige $\alpha < \frac{1}{\lambda_1}$ con λ_1 valor propio dominante de A

  $\Rightarrow$ asegura que $I - \alpha A^t$ sea invertible

- Se puede ver que X cumple que $X = \mathbf{1} + \alpha A\mathbf{1} + \alpha^2 A^2 \mathbf{1} + \cdots$ ...

  $\Rightarrow$ Si $\alpha \to 0$ entonces tenemos la centralidad por grado

  $\Rightarrow$ Si $\alpha \to \frac{1}{\lambda_1}$ entonces tenemos la centralidad por vector propio

Centralidad Katz corregida

- Uno entre muchos... un vértice con centralidad Katz alta, le hereda su importancia a todos sus “out-vecinos”

  $\Rightarrow$ Amazon es importante y me linkea ¿soy importante?

- Idea: importancia es proporcional a la importancia de mis (in)-vecinos dividido por su grado saliente

$$x_i = \alpha \sum_j A^t_{ij} \frac{x_j}{d^{out}_j} + \beta$$

- Si D es diagonal con $d_{ii} = d^{out}_i$:

$$X = \alpha A^t D^{-1} X + \beta \mathbf{1}$$

$\Rightarrow$ $d^{out}_j = 0$ se cambia por $d^{out}_j = 1$, definiendo $d_{ii} = \max\{d^{out}_i, 1\}$

Page Rank

- Eligiendo $\beta = 1$:

$$X = (I - \alpha A^t D^{-1})^{-1}\mathbf{1}$$

- Famoso Page Rank que usa(ba) Google(Brin-Page, 1998) con $\beta = \frac{1-\alpha}{N_v}$
- ¿Cómo se elige α?

  $\Rightarrow$ siguiendo las mismas ideas que antes: $\alpha < \frac{1}{\lambda_1}$, con λ_1 valor propio dominante de $A^t D^{-1}$ o AD^{-1} (se resumen en $\alpha < 1$)

  $\Rightarrow$ Google usa(ba) $\alpha = 0{,}85$ aunque no se sabe porqué...

Resumen

	Con importancia de regalo	Sin importancia de regalo
Dividir por grado saliente	$X = (I - \alpha A^t D^{-1})^{-1}\mathbf{1}$ Page Rank	$X = AD^{-1}\mathbf{1}$ Centralidad de grado
Sin dividir	$X = (I - \alpha A^t)^{-1}\mathbf{1}$ Centralidad Katz	$X = k^{-1}AX$ Centralidad de vector propio

- Centralidad de grados: todos los vecinos tienen la misma importancia
- Centralidad de vector propio (CVP): la importancia de un vértice es proporcional a la importancia de sus vecinos
- Centralidad Katz: la CVP no funciona para dirigidos. Se arregla dando importancia de regalo a todos los vértices
- Page Rank: la importancia de un vértice es proporcional a la importancia de sus vecinos ponderado por su grado saliente

Hubs and authorities

- Problema de ranking de páginas web o redes de citas: dos tipos de nodos importantes hubs/centros and authorities/autoridades

  $\Rightarrow$ importancia de apuntar hacia nodos importantes

- Autoridades son nodos (páginas) que contienen información últil o relevante
  - páginas de diarios importantes, página de Bedelía, Gallito Luis
- Centros son los expertos que nos dicen dónde encontrar a las autoridades
  - Listado de diario, Instagram de la facultad, Páginas amarillas, Paper tutorial
- Autoridades y centros se refuerzan mutuamente

  $\Rightarrow$ Un buen centro apunta a varias buenas autoridades

  $\Rightarrow$ Una buena autoridad apunta a varios buenos centros

Ranking de autoridades y centros

- Algoritmo Hyperlink-Induced Topic Search (HITS) [Kleinberg'98]
- Cada nodo v tiene una medida de centro h_v y una medidad de autoridad a_v

  $\Rightarrow$ dos vectores definidos en la red $\mathbf{h} = [h_1, \ldots, h_{N_v}]^\top$, $\mathbf{a} = [a_1, \ldots, a_{N_v}]^\top$

  $\Rightarrow$ se puede tener centralidad de autoridad nula pero centralidad de centro no-nula

Regla de actualización de autoridades:

$$a_v(k) = \sum_{(u,v)\in E} h_u(k-1), \text{ for all } v \in V \Leftrightarrow \mathbf{a}(k) = \mathbf{A}^\top \mathbf{h}(k-1)$$

Regla de actualización de centros:

$$h_v(k) = \sum_{(v,u)\in E} a_u(k), \text{ for all } v \in V \Leftrightarrow \mathbf{h}(k) = \mathbf{A}\mathbf{a}(k)$$

- Iniciar con $\mathbf{h}(0) = \mathbf{1}/\sqrt{N_v}$, y normalizar $\mathbf{a}(k)$ and $\mathbf{h}(k)$ en cada iteración

Valores límites

- Se define el ranking de centros y autoridades como los valores límites

$$\mathbf{a} := \lim_{k\to\infty} \mathbf{a}(k), \quad \mathbf{h} := \lim_{k\to\infty} \mathbf{h}(k)$$

- Dada las reglas de actualización se tiene que $k = 0, 1, \ldots$

$$\mathbf{a}(k+1) = \frac{\mathbf{A}^\top \mathbf{A}\mathbf{a}(k)}{\|\mathbf{A}^\top \mathbf{A}\mathbf{a}(k)\|}, \quad \mathbf{h}(k+1) = \frac{\mathbf{A}\mathbf{A}^\top \mathbf{h}(k)}{\|\mathbf{A}\mathbf{A}^\top \mathbf{h}(k)\|}$$

- Nuevamente por método de las potencias converge a los vectores propios de $\mathbf{A}^\top\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$

$$\mathbf{A}^\top \mathbf{A}\mathbf{a} = \alpha^{-1}\mathbf{a}, \quad \mathbf{A}\mathbf{A}^\top \mathbf{h} = \alpha^{-1}\mathbf{h}$$

$\Rightarrow$ El ranking de autoridades y centros son centralidades de vector propio

Closeness Centrality

- Idea: un vértice es “central” si está cerca de muchos otros vértices

Definición (Distancia)

Sea $G(V, E)$ grafo simple, no dirigido. La distancia $d(u, v)$ entre dos vértices u y v es el largo del camino más corto entre u y v. Se suele referir como distancia geodésica.

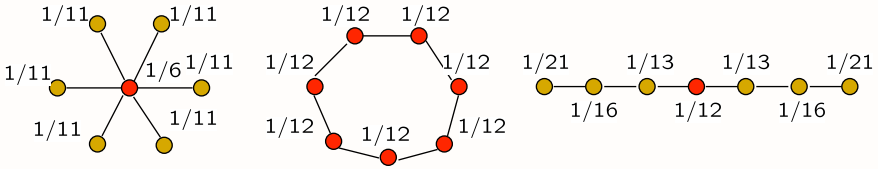
Definición (Closeness centrality)

La centralidad de cercanía de v se define como

$$c(v) = \frac{1}{\sum_{u \in V} d(u, v)}$$

- Se puede interpretar $v^* = \arg\max_v c_(v)$ como el vértice más alcanzable en G

Ejemplos



- Izquierda: centralidad de grado, de vector propio y de cercanía coinciden en el ranking
- Centro: centralidad de grado, de vector propio y de cercanía coinciden en el ranking
- Derecha: centralidad de vector propio y cercanía coinciden en el ranking pero diferente de grado

Centralidad de cercanía

- Ejemplo: red de actores (Christopher Lee) y el Bacon number
- Para facilitar la comparación se normaliza para que el rango sea $[0, 1]$

$$c(v) = \frac{N_v - 1}{\sum_{u \in V} d(u, v)}$$

- Nodos muy conectados tienden a reducir distancias
- Limitación: sensible, valores se mueven en un rango pequeño

  $\Rightarrow$ difícil de discriminar nodos centrales de no tan centrales
- Limitación : asume G conexo, sino $c(v) = 0$ for all $v \in V$

  $\Rightarrow$ se puede calcular ranking por componente, pero el orden de la componente influye

Centralidad de cercanía

Definición (alternativa)

Uan definición alternativa es la media armónica de las distancias:

$$c'(v) = \frac{1}{N_v - 1} \sum_{u \neq v} \frac{1}{d(u, v)}$$

- Da importancia a los nodos más cercanos (importa más qué tan cerca estoy de los más cercanos)

  $\Rightarrow$ Pero no se usa...

Betweenness centrality

- Idea: Un vértice es “central” si está en el camino entre muchos pares de vértices

Definición (Betweenness centrality)

La centralidad de entremedio de un vértice v está dada por

$$c(v) = \sum_{s \neq t \neq v \in V} \frac{n_{st}^v}{g_{st}} \quad \text{donde}$$

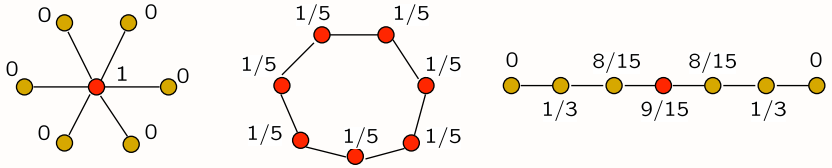
- g_{st} número total de caminos más cortos entre s y t
- n_{st}^v es el número de caminos más cortos entre s y t que pasan por v

- Se puede Interpret $v^* = \arg\max_v c(v)$ como el controlador del flujo de información

Betweenness centrality

- ¿Ejemplo de vértice con grado bajo pero betweenness alta?
- Ejemplo de antes (atención! los números correspoden a normalizar por la cantidad de pares de origen-destino $(N_v - 1)(N_v - 2)$)

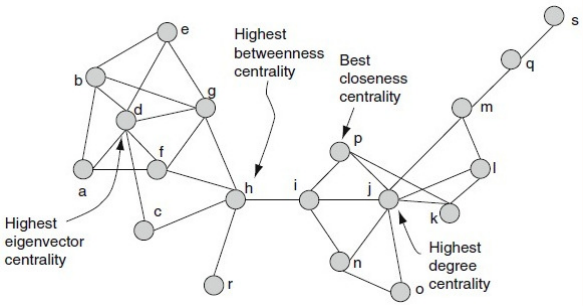
  $\Rightarrow$ el ranking sin normalizar es el mismo



- Red de actores: Fernando Rey (español/inglés) y segundo Christopher Lee (también tiene poco rango)
- Complejidad computacional

Comparación de métricas de centralidad

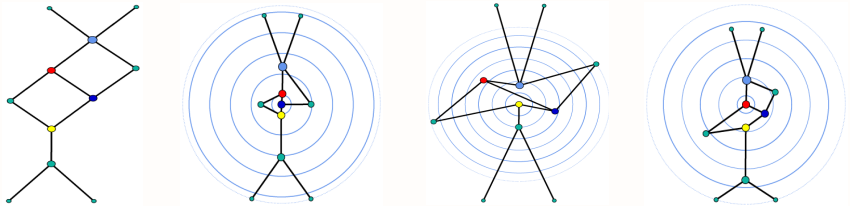
- ¿Cuál vértice es el más central? Depende del contexto



- Cada medida identifica un vértice diferente como el más central
- Ninguno está 'mal', las medidas apuntan a diferentes nociones de importancia

Comparación de métricas de centralidad

- ¿Cuál vértice es el más central? Depende del contexto



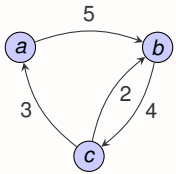
Closeness Betweenness Eigenvector

- Los verdes chiquitos parecen ser más periféricos

  $\Rightarrow$ Es menos claro como se comparan el amarillo, azul y rojo

Definiciones para digrafos con pesos

- Grafos con pesos y dirigidos graphs $G(V, E, W)$
  - $\Rightarrow$ Función $W : E \to \mathbb{R}^+$ de pesos en cada arista
  - $\Rightarrow$ Matriz de adyacencia no binaria ni simétrica pero con entradas positivas



- Camino $P(u, v)$ es una secuencia ordenada de vértices y aristas entre u to v
- Largo del camino es la suma de los pesos de los arcos del camino
- Largo del camino más corto $s_G(u, v)$ de u a v

$$s_G(u, v) := \min_{P(u,v)} \sum_{i=0}^{\ell-1} W(u_i, u_{i+1})$$

Centralidades para digrafos con pesos

- Centralidad de grado: suma de los peso de los arcos entrantes

$$c(v) := \sum_{u|(u,v)\in E} W(u,v)$$

- Centralidad de vector propio: vector propio asociado al valor propio dominante de A^t (matriz positiva, vale PF)
- Closeness y betweenness: basta cambiar la definición de largo de camino (ahora es la suma de los pesos en el camino)

Estabilidad de medidas de centralidad

- Espacio de grafos $\mathcal{G}_{(V,E)}$ con (V, E) conjunto de vértices y aristas
- Se define la métrica $d_{(V,E)}(G, H) : \mathcal{G}_{(V,E)} \times \mathcal{G}_{(V,E)} \to \mathbb{R}_+$

$$d_{(V,E)}(G, H) := \sum_{e \in E} |W_G(e) - W_H(e)|$$

- **Definición:** Una medida de centralidad se dice $c(\cdot)$ is *estable* si para todo vértice en $v \in V$ en dos grafos cualesquiera $G, H \in \mathcal{G}_{(V,E)}$, se cumple

$$\left| c^G(v) - c^H(v) \right| \leq K_G \, d_{(V,E)}(G, H)$$

  - $\Rightarrow$ K_G es una constante que solo depende de G
  - $\Rightarrow$ Relación con la continuidad Lipschitz en el espacio $\mathcal{G}_{(V,E)}$
  - $\Rightarrow$ Independente de la definición de la métrica (equivalencia de normas)

Estabilidad de medidas de centralidad

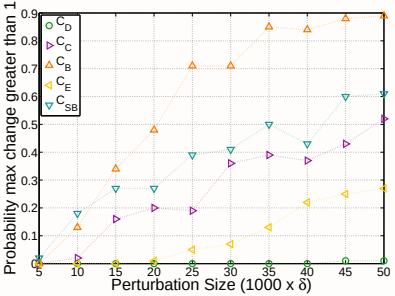
- La importancia de los vértices tiene que ser robusta frente a pequeñas perturbaciones del grafo
- La centralidad por grado es estable
- Las centralidades de vector propio son estables
- La centralidad de cercanía (closeness) es estable
- La centralidad betweenneess no es estable... se puede modificar para que lo sea

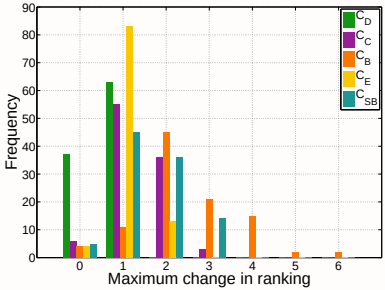
  $\Rightarrow$ S. Segarra and A. Ribeiro, “Stability and continuity of centrality measures in weighted graphs,” *IEEE Trans. Signal Process.*, 2015.

Ejemplo en el grafo de aeropuertos

- Grafo real basado en el tráfico aéreo entre aeropuertos más populares de EEUU
  - $\Rightarrow$ Vértices $N_v = 25$,
  - $\Rightarrow$ Peso de las aristas: número de pasajeros anuales entre ellos
- Perturbación multiplicar los pesos por números aleatorios $\delta \sim \mathcal{U}(0{,}99, 1{,}01)$
- Calculamos probabilidad de observar cambios en el ranking
- Grafica del histograma de cambio máximo

Ejemplo en el grafo de aeropuertos





- La centralidad betweenness es la que muestra las mayores variaciones

Cohesión de la red

- Otro aspecto importante en análisis de redes refiere a la cohesión de la red

Ejemplo

- ¿Mis amigos son amigos entre sí?
- ¿Qué grupo de proteínas en una célula trabajan más estrechamente?
- ¿Se separa la estructura de las páginas web en relación con el contenido?
- ¿Qué parte de la topología de Internet constituye un "backbone' ?

- Las definiciones de cohesión en una red dependen del contexto

  $\Rightarrow$ Escala local (e.g., tríos) a global (e.g., componentes gigantes)

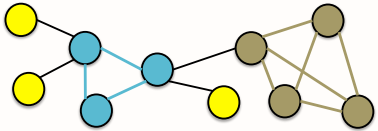
  $\Rightarrow$ Explicítamente específicadas (e.g., cliques) or implícitamente (e.g., clusters)

Subgrupos cohesivos

- Subgrupos cohesivos definidos por analistas de redes sociales como:
  'Actores conectados a través de relaciones densas, dirigidas y recíprocas'
- Permitir compartir información, crear solidaridad, coordinar acciones colectivas
  Ejemplos: clubes deportivos, cultos religiosos, células terroristas, organizaciones estudiantiles, ...
- Propiedades deseables de un subgrupo cohesivo
  - $\Rightarrow$ Familiaridad (grado);
  - $\Rightarrow$ Alcanzabilidad (distancia);
  - $\Rightarrow$ Robustez (conectividad); y
  - $\Rightarrow$ Densidad (densidad de aristas)
- Bastante natural pensar en cliques, i.e., subgrafos completos del grafo asociado a la red

Densidad local y cliques

- Cliques grandes son raros; basta que falte una arista



- Una condición suficiente para la existencia de un n-clique es

$$N_e > \frac{N_v^2}{2}\frac{(n-2)}{(n-1)}, \text{ mientras que los grafos } \textit{sparse} \text{ tienen } N_e = O(N_v)$$

- La complejidad de algoritmos relativos a cliques varía ampliamente
  - ¿Es $U \subseteq V$ un clique? ¿Es maximal? $O(N_v + N_e)$
  - Identificar todos los triángulos en G: $O(N_v^3)$ ($O(N_v^{\sqrt{2}})$ para grafos *sparse*)
  - ¿ G tiene un clique maximal de tamaño $\geq n$? NP-complete

Coeficiente de clustering

- ¿Que fracción de los vecinos de v están a su vez conectados entre sí?

Definición

El clustering coefficient $\mathrm{cl}(v)$ de $v \in V$ está dado por

$$\mathrm{cl}(v) = \frac{2|E_v|}{d_v(d_v - 1)} \in [0, 1]$$

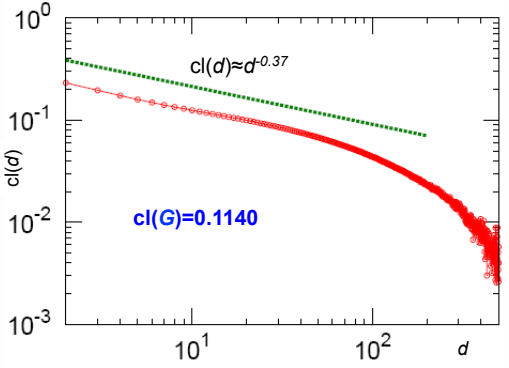
donde $|E_v|$ es el número de aristas entre los vecinos de v.

cl(v)=0 cl(v)=1/3 cl(v)=1

- Es una indicación de qué tanto las aristas forman “cluster”
- El clustering coefficient global es el promedio $\mathrm{cl}(G) = \frac{1}{N_v} \sum_{v \in V} \mathrm{cl}(v)$

Ejemplo: MSN social network

- MSN social network: $N_v \approx 180M$, $N_e \approx 1{,}3B$ [Leskovec et al'06]



- Clustering coefficient global $\text{cl}(G) = 0{,}1140$ es grande
- Comparado con aristas al azar e igual probabilidad p

$$\text{cl}(G_{n,p}) = \Pr[\text{arista cierre un triángulo}] = p = \frac{c}{n-1} \to 0$$

Centralidad para grupos de vértices

- Captura la importancia de un subgrupo de vértices [Everett et al'99]

  $\Rightarrow$ ¿Los ingenieros son más populares que los contadores en una cierta organización?

  $\Rightarrow$ ¿Cómo elegimos los representantes con mayor influencia?

- Generalización de las centralidades de vértices

Definición (Centralidad de grado grupal)

Sea el subgrafo $G'(V', E')$ inducido por el subconjunto de vértices V' y sea $U_{V'} \subset V \setminus V'$ conjunto de vértices fuera de V' con aristas hacia vértices en V'. Se define la centralidad de grado de V' como:

$$d_{V'} = |U_{V'}|$$

- Número de vértices fuera del grupo pero conectados con vértices del grupo

Centralidad para grupos de vértices

- **Def:** Distancia de $v \in V$ a un grupo de vértices $V' \subset V$ es

$$d_*(v, V') = \min_{u \in V'} d(u, v)$$

- Centralidad de cercanía grupal de un subconjunto de vértices V'

$$c_{Cl}(V') = \frac{1}{\sum_{u \in V \setminus V'} d_*(u, V')}$$

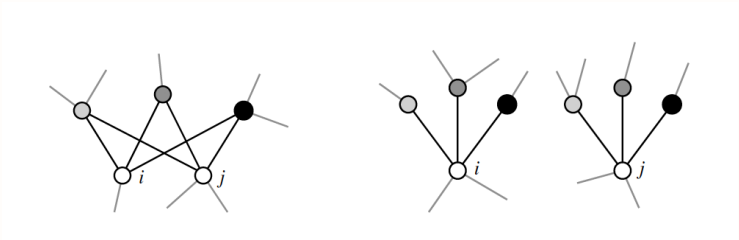
- Centralidd betweenness grupal de un subconjunto de vértices V'

$$c_{Be}(V') = \sum_{s \neq t \in V \setminus V'} \frac{\sigma(s, t|V')}{\sigma(s, t)}$$

  - $\sigma(s, t)$ número total de $s - t$ caminos más cortos $(s, t \in V \setminus V')$
  - $\sigma(s, t|V')$número de $s - t$ caminos más cortos que pasan por $v \in V'$

Similaridad

- Usando la información contenida en la estructura de la red ¿Es posible definir similaridad entre nodos?
- Equivalencia estructural: dos nodos son estructuralmente equivalentes si comparten mucho de su red de vecinos
- Equivalencia regular: no necesariamente comparten los mismos vecinos pero tienen vecinos que son similares entre ellos



Equivalencia estructural

- Lo más sencillo es la cantidad de vecinos en común:

$$\eta_{ij} = \sum_k A_{ik} A_{kj} = (A^2)_{ij}$$

- Pero si es poco o mucho depende de los grados de cada nodo...

Definición (Similitud del coseno)

Se define la similitud entre los vértices i y j como el ángulo entre las filas i y j de la matriz de adyacencia:

$$\sigma_{ij} = \frac{\eta_{ij}}{\sqrt{d_i}\sqrt{d_j}}$$

- Si $d_i = 0$ o $d_j = 0$, se define $\sigma_{ij} = 0$
- Ejemplo anterior: $\sigma_{ij} = \frac{3}{2\sqrt{5}} = 0,671$.

Equivalencia estructural

- Propiedades de la similitud del coseno:
  - $\Rightarrow$ Es simétrica
  - $\Rightarrow 0 \leq \sigma_{ij} \leq 1$ para todo $i, j \in V$ (fácil comparar)
  - $\Rightarrow \sigma_{ii} = 1$ (un nodo es similar a sí mismo)

Definición (Coeficiente Jaccard)

Para comparar la cantidad de vecinos en común, con la cantidad de vecinos distintos (los comunes cuentan una sola vez)

$$J_{ij} = \frac{\eta_{ij}}{d_i + d_j - \eta_{ij}}$$

- Propiedades del coeficiente de Jaccard: es simétrico y $0 \leq J_{ij} \leq 1$
  - $\Rightarrow J_{ii} = 1$ (un nodo es similar a sí mismo)
  - $\Rightarrow J_{ij} = 0 \Leftrightarrow \sigma_{ij} = 0$ y $J_{ij} = 1 \Leftrightarrow \sigma_{ij} = 1$

Equivalencia estructural

Definición (Correlación de Pearson)

Se define la similitud entre los vértices i y j como el coeficiente de correlación de Pearson entre las filas i y j de la matriz de adyacencia

$$r_{ij} = \mathrm{corr}(A_{i\cdot}, A_{j\cdot}) = \frac{\sum_k (A_{ik} - \bar{A}_i) \sum_k (A_{jk} - \bar{A}_j)}{\sqrt{\sum_k (A_{ik} - \bar{A}_i)^2} \sqrt{\sum_k (A_{jk} - \bar{A}_j)^2}}$$

- Propiedades del coeficiente de correlación:
  - $\Rightarrow$ Es simétrico y $-1 \leq r_{ij} \leq 1$ para todo $i, j \in V$
  - $\Rightarrow$ $r_{ii} = 1$ (un nodo es similar a sí mismo)
  - $\Rightarrow$ El coeficiente de correlación, es el ángulo entre los vectores normalizados (centrados y con varianza 1)

Equivalencia estructural

Definición (Distancia de Hamming)

Se define la disimilitud entre los vértices i y j simplemente como el número de vecinos que no son vecinos en común (vecinos de uno pero no del otro).

$$h_{ij} = \sum_k (A_{ik} - A_{jk})^2$$

- Propiedades de la distancia de Hamming:
  - $\Rightarrow$ es simétrica, $0 \leq h_{ij} \leq \max\{d_i, d_j\}$
  - $\Rightarrow$ $h_{ii} = 0$ (un nodo no es (dis)similar a sí mismo)
- Ejemplo: $\sigma_{ij} = 0,671$, $r_{ij} = -0,316$, $J_{ij} = \frac{3}{6} = 0,5$, $h_{ij} = 3$
  - $\Rightarrow$ Más en Wasserman and Faust "*Social Network Analysis*', Cambridge University Press (1994).

Equivalencia regular

Definición (Equivalencia regular)

Se define una medida de similitud σ_{ij} tal que i y j tienen similitud alta si tienen vecinos k y l también con similitud alta

$$\sigma_{ij} = \alpha \sum_{kl} A_{ik} A_{jl} \sigma_{kl} \quad \text{con } \alpha \text{ constante de proporcionalidad}$$

- En matrices $\sigma = \alpha A \sigma A$
- Problema: σ_{ii} no es necesariamente alta, se agrega un término diagonal:

$$\sigma_{ij} = \alpha \sum_{kl} A_{ik} A_{jl} \sigma_{kl} + \delta_{ij}$$

$$\sigma = \alpha A \sigma A + I$$

- Se puede ver que solo depende de los caminos de largo par entre i y j

Equivalencia regular

Definición (Similitud de Katz)

Se define una medida de similitud σ_{ij} tal que i y j tienen similitud alta si i tiene un vecinos k que tiene similitud alta con j

$$\sigma_{ij} = \alpha \sum_{kl} A_{ik}\sigma_{kj} + \delta_{ij} \quad \text{con } \alpha \text{ constante de proporcionalidad}$$

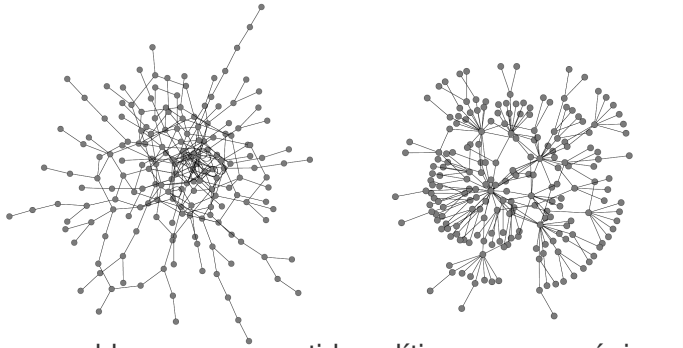
- En matrices $\sigma = \alpha A\sigma + I$
- Se puede ver que $\sigma_{ij} = \sum_m \alpha^m (A^m)_{ij}$ vinculado a caminos de largo m
- La elección de α es la usual $\alpha < \frac{1}{\lambda_1}$ valor propio dominante de A
- Similitud Page Rank: dividir por el grado

$$\sigma_{ij} = \frac{1}{d_i}\left(\alpha \sum_{kl} A_{ik}\sigma_{kj} + \delta_{ij}\right)$$

Assortative mixing

- Personas tienen una fuerte tendencia a asociarse con “iguales”

  $\Rightarrow$ esta tendencia se llama homofilia o assortative mixing



- Ej: liceales por raza, bloggers por partido político, papers, páginas web . . .
- Al contratio disassortative mixing es asociación por el contrario e.g., redes de encuentros sexuales

Cuantificando la 'asortatividad'

- Supongamos que cada vértice i es de tipo g_i donde las características son categóricas (sin orden), e.g., hombre/mujer
- Número de aristas (i, j) tal que i, j son del mismo tipo:

$$\sum_{(i,j)\in E} \delta_{g_i g_j} = \frac{1}{2} \sum_{i,j} A_{ij} \delta_{g_i g_j} \quad \text{con} \quad \delta_{g_i g_j} = \begin{cases} 1 & \text{si } g_i = g_j \\ 0 & \text{otro caso} \end{cases}$$

- Número esperado de aristas entre dos vértices del mismo tipo si las aristas fueran colocadas al azar (respetando los grados):

$$\frac{1}{2} \sum_{i,j} \frac{d_i d_j}{2N_e - 1} \delta_{g_i g_j} \sim \frac{1}{2} \sum_{i,j} \frac{d_i d_j}{2N_e} \delta_{g_i g_j}$$

‘Asortatividad’ por características sin orden

- Se define la modularidad Q como la diferencia entre el número de aristas presente y la esperada ponderada por la cantidad de aristas total

$$Q = \frac{1}{2N_e} \sum_{ij} \left(A_{ij} - \frac{d_i d_j}{2N_e} \right) \delta_{g_i g_j} < 1$$

  $\Rightarrow Q > 0$ es indicación por “iguales”, assorative mixing

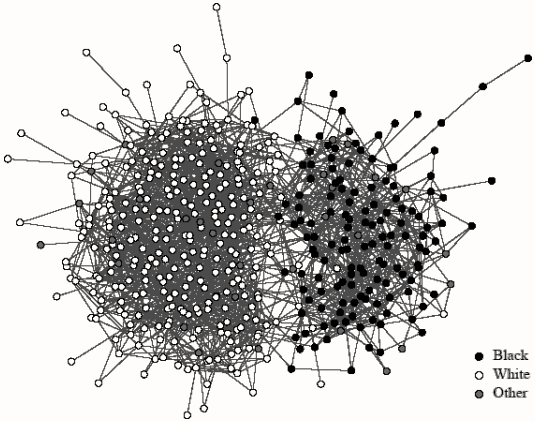
  $\Rightarrow Q < 0$ es indicación por “contrario”, disassortative mixing

- Otra forma de calcular la modularidad que puede ser útil en algunos formatos de datos:

$$Q = \sum_r (e_r - a_r^2)$$

  - e_r proporción de aristas entre vértices de tipo r
  - a_r proporción de arcos que apuntan a nodos de tipo r

‘Asortatividad’ por características sin orden



- Red de amigos en un liceo de USA
- Modularidad $Q = 0{,}305$
- Indicativo de assortative mixing por raza

‘Asortatividad’ por características con orden

- Características como edad, ingresos, nivel educativo tiene orden
- Podemos indicar cuándo las características ‘están cerca’ (ej:cumplen la misma semana)
- ¿Cómo calcular la modularidad en este caso?

  $\Rightarrow$ definir grupos por valor de la característica: solo considero los idénticos

  $\Rightarrow$ definir alguna partición (joven/adulto/anciano): se pierde capacidad de diferenciar
- No parece conveniente usar la misma métrica
- Para cada vértice i, sea x_i su característica (valor numérico)

  $\Rightarrow$ para cada arista (i, j) consideramos los pares (x_i, x_j) asociados

‘Asortatividad’ por características con orden

- Calculamos la covarianza empírica de los (x_i, x_j) sobre todas las aristas:

$$\operatorname{cov}(x_i, x_j) = \frac{1}{2N_e} \sum_{ij} \left(A_{ij} - \frac{d_i d_j}{2N_e} \right) x_i x_j$$

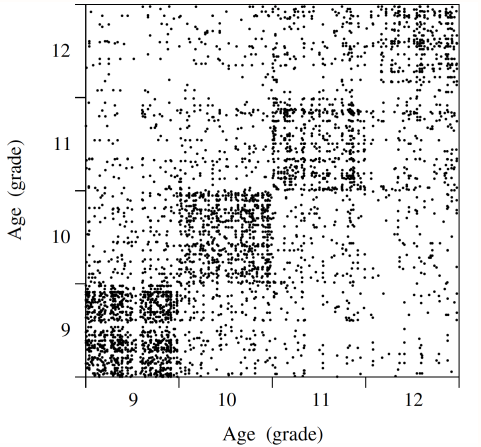
  $\Rightarrow$ observar que solo cambia $\delta_{g_i g_j}$ por $x_i x_j$

- $\operatorname{cov}(x_i, x_j) > 0$: assortative mixing y $\operatorname{cov}(x_i, x_j) < 0$: disassortative mixing
- Assortative coefficient: normalizar para que valga 1 en el caso de assortative mixing perfecto:

$$r = \frac{\sum_{ij} \left(A_{ij} - d_i d_j / 2N_e \right) x_i x_j}{\sum_{ij} \left(A_{ij} \delta_{ij} - d_i d_j / 2N_e \right) x_i x_j}$$

- Aunque no es evidente, r es el coeficiente de correlación usual

‘Asortatividad’ por características con orden



- Mismo ejemplo que antes pero con la edad (edad equiparada a grado)
- Coeficiente de assortatividad $r = 0{,}616$
- Indica un alto nivel de assortative mixing por edad

‘Asortatividad’ por grado

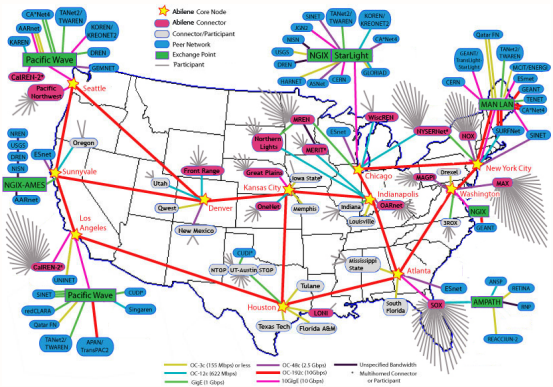
- Assortative mixing por grado: nodos de grado alto se juntan con nodos de grado alto
- Dissasortative mixing por grado: nodos de grado alto se juntan con nodos de grado bajo
- Basta cambiar x_i por d_i en la fórmula anterior:

$$r = \frac{\sum_{ij} \left(A_{ij} - d_i d_j / 2N_e\right) d_i d_j}{\sum_{ij} \left(A_{ij}\delta_{ij} - d_i d_j / 2N_e\right) d_i d_j}$$

- Solo necesitamos la matriz de adyacencia

Ejemplo: Abilene network

- Red Abilene entre universidades en USA y laboratorios de investigación
  - Grupos categóricos: ‘Core’ , ‘Connector’, ‘Exchange points’ , etc



- Estructura jerárquica nos sugiere disassortative mixing

Disassortative mixing en Abilene

- Tabla con el conteo de aristas entre categorías

	Core	Exchange	Peer	Conn.	Part.	Conn./Part.
Core	14	6	5	17	0	16
Exchange	6	1	46	2	0	0
Peer	5	46	0	0	0	1
Conn.	17	2	0	0	203	0
Part.	0	0	0	203	0	34
Conn./Part.	16	0	1	34	34	0

- Modularidad $Q = -0{,}3162$

  $\Rightarrow$ Fuerte soporte a la sugerencia de disassortative mixing