Caminatas

Caminos

Dada una red, distintos tipos de flujo pueden ser de interés

Imaginemos flujo de objetos (items indivisibles que están en un único lugar a un dado tiempo). Su **difusión es por transferencia**.
Puede haber casos en que:

- se mueven por el grafo **sin restricciones** acerca de repetir enlaces o nodos ya visitados (e.g. billetes, libros, …).

- se mueven por el grafo **sin repetir enlaces** (i.e. ropa usada, *re-gifting*)

- Correo: se mueven por el grafo desde un nodo origen hacia uno destino minimizando distancia recorrida.

Caminos

camino (*walk*): secuencia de vertices tales que vértices consecutivos están conectados (en grafos dirigidos la secuencia debe respetar las orientaciones de los enlaces)

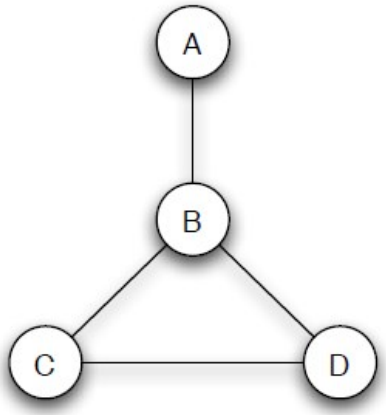
sendero (*trail*): es un camino que no repite enlaces

camino simple (*path*):es un camino que no repite vértices

ciclo: camino de más de dos vértices que no repite elementos, salvo el primero y último

longitud de un camino C: número de pasos que contiene desde su comienzo hasta su final (Para redes pesadas, **longitud** asociado a un camino: suma de los pesos de los enlaces involucrados)

camino geodésico : camino de menor longitud que une un par dado de vértices.

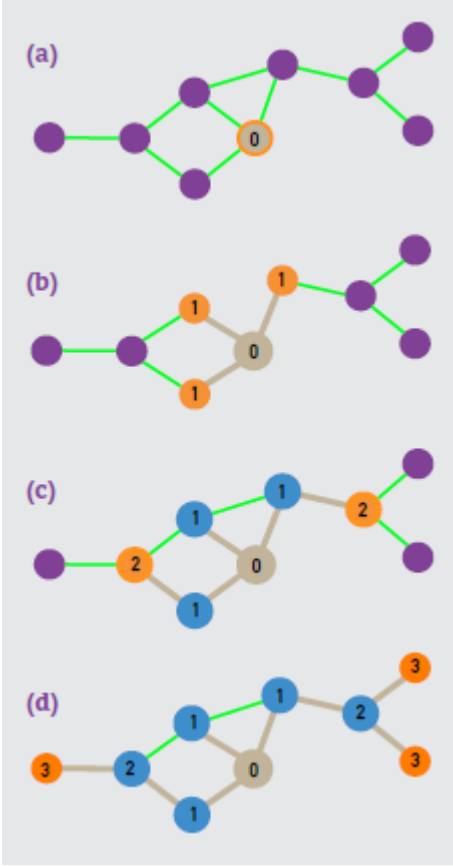


c_{walk}={A,B,D,B,C}
c_{trail}={A,B,D,C,B}
c_{path}={A,B,C}
c_{ciclo}={B,C,D,B}

Longitud de caminos

Buscando a lo ancho…*Breath-First Search*

I. Empezar en nodo i, etiquetarlo como ´0´
II. Encontrar sus primeros vecinos, etiquetarlos como ´1´ y ponerlos *en espera* (en cola de procesamiento)
III. Tomar el primer nodo de etiqueta ´e´ de la cola (*e=1*, la primera vez).
    a. Encontrar sus nodos adyacentes sin etiqueta
    b. Etiquetarlos como *e+1* y ponerlos en la cola de procesamiento
IV. Repetir III hasta encontrar el nodo *j* buscado, o que no queden mas nodos en la cola.
V. La distancia entre *i* y *j* es la etiqueta de *j*. Si no tiene ninguna $d_{ij} = \infty$



Caminos

- Cuántos caminos de una dada longitud λ existen en una red?

$A_{ij} \neq 0$ sii existe un enlace que vaya de *nodo-j* hacia *nodo-i*

$A_{ik}A_{kj} \neq 0$ sii existe un camino de *nodo-j* hacia *nodo-i*, pasando por *nodo-k*

$A_{ik}A_{kl}A_{lj} \neq 0$ sii existe un camino de *nodo-j* hacia *nodo-i*, pasando por *nodo-k* y *nodo-l*

$$NC^{(2)}{}_{ij} = \sum_{k=1}^{N} A_{ik}A_{kj} = [\boldsymbol{A}^2]_{ij}$$

$$NC^{(\lambda)}{}_{ij} = \sum_{k,l,\ldots,s=1}^{N} \overbrace{A_{ik}A_{kl}\ldots A_{sj}}^{\lambda} = \left[\boldsymbol{A}^{\lambda}\right]_{ij}$$

- Cuántos **ciclos** existen de una dada longitud λ existen en la red?

$$NC^{(\lambda)}{}_{ii} = \left[\boldsymbol{A}^{\lambda}\right]_{ii} \qquad NC^{(\lambda)} = \sum_{i=1}^{N}\left[\boldsymbol{A}^{\lambda}\right]_{ii} = Tr(\boldsymbol{A}^{\lambda})$$

←ojo

Caminos

- Cómo saber si mi grafo es acíclico?

Un grafo de matriz de adyacencia ***A*** es acíclico ↔ **todos** los autovalores de ***A*** son nulos

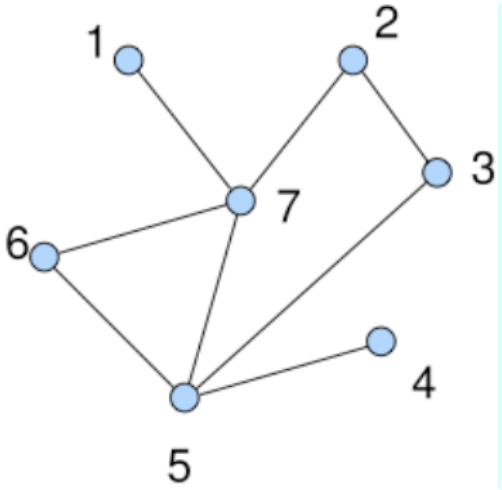
→

- si G es acíclico ***A*** puede escribirse como una matriz triangular superior (mts)
- una mts tiene diagonal cero… entonces todos sus autov. son nulos

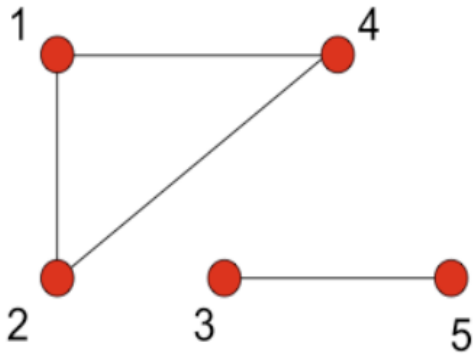
←

$$NC^{(\lambda)} = \sum_{i=1}^{N} \left[\boldsymbol{A}^{\lambda}\right]_{ii} = Tr\left(\boldsymbol{A}^{\lambda}\right) = \sum_{i=1}^{N} (\alpha_i)^{\lambda} = 0$$

Componentes de un grafo



Grafo conexo

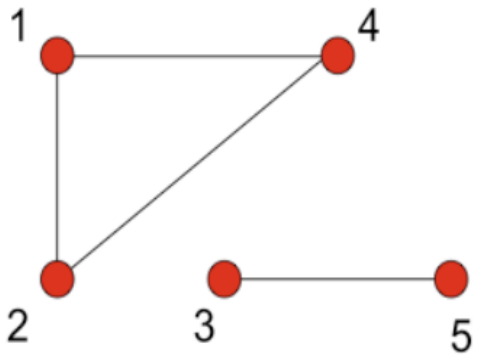


Grafo disconexo de dos componentes

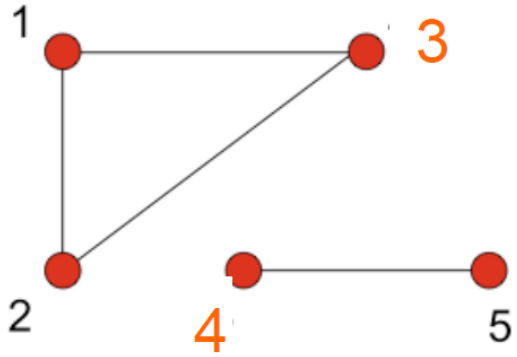
Componente de un grafo subconjunto *maximal* de vértices tal que existe un camino entre cualesquiera dos de sus elementos

Componentes de un grafo

$$A = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix}$$

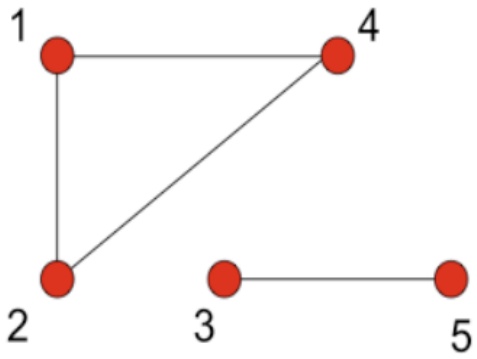


$$A' = \begin{array}{c} \begin{matrix} 1 & 2 & 4 & 3 & 5 \end{matrix} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 4 \\ 3 \\ 5 \end{matrix}$$



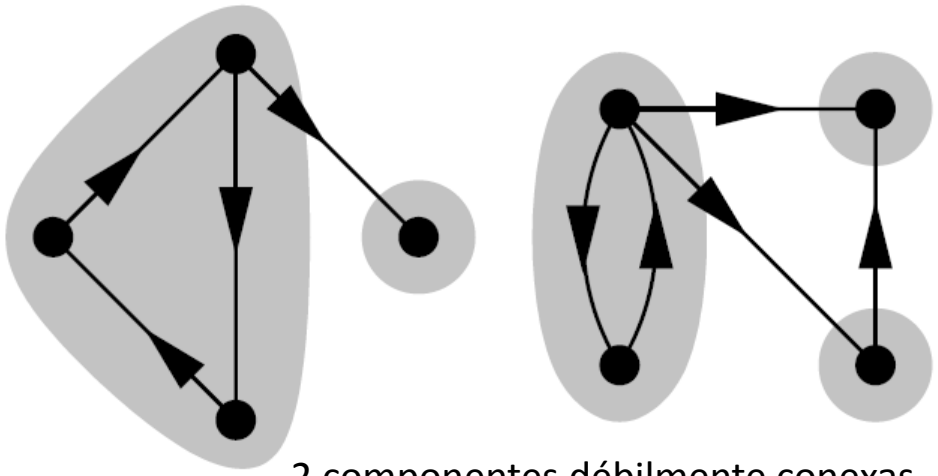
La matriz de un grafo disconexo se puede llevar, mediante permutaciones de filas y columnas, a una **matriz diagonal en bloques**

Buscando componentes

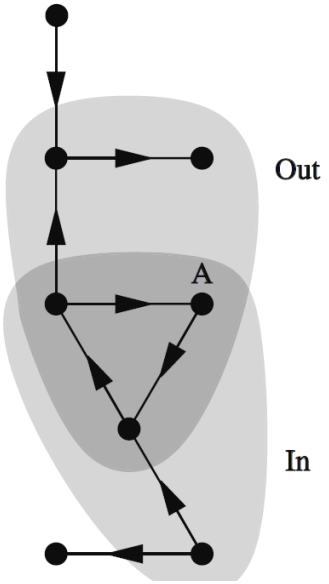


I. Realizar un BFS desde un nodo-i. Etiquetar a **todos** los nodos alcanzados como *c=1*
II. Si el número total de nodos etiquetados es igual al número de nodos de la red entonces *stop*. Sino continuar a III
III. Elegir un nodo no etiquetado *j*. Hacer $c \leftarrow c + 1$. Realizar un BFS desde el nodo-j y etiquetar a todos los nodos alcanzados como *c.* Repetir II
IV. El número de componentes resulta *c*

Componentes de un **grafo dirigido**



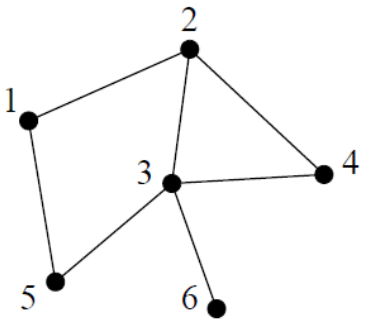
2 componentes débilmente conexas
5 componentes fuertemente conexas



- componente **débilmente conexa**: componente conexa del grafo no-dirigido asociado
- componente **fuertemente conexa**: subconjunto maximal de nodos tal que existen caminos dirigidos entre cualquier par ordenado del conjunto.
- **out-component inducido por un nodo-i**: subconjunto de nodos alcanzables desde nodo-i a través de caminos dirigidos del grafo
- **in-component inducido por un nodo-i**: subconjunto de nodos desde los que se alcanza al nodo-i a través de algun caminos dirigido del grafo

Caminatas al azar

- Imaginemos un caminante (Juan) severamente alcoholizado caminando por una red (simple y conexa / fuertemente conexa).
- Consideremos el tiempo de manera discreta.
- A *t=0* Juan se encuentra en el nodo-*k*
- Al pasar un intervalo *Δt* Juan avanza desde el nodo donde se encuentra actualmente hacia alguno de sus vecinos



¿Cuál es la probabilidad ,$p_i(t)$, que a tiempo *t* Juan se encuentre en el nodo *i ?*

Por qué tiene sentido esta pregunta?

- Inicialmente Juan está en el nodo-*s.* El **vector** de probabilidades ***p(t=0)*** es

$$p_i(t=0) = \delta_{si} = \begin{cases} 1 \; si \;\; i = s \\ 0 \; si \; i \neq s \end{cases}$$

- Por cómo es el proceso de desplazamiento ***p(t+1)*** va a ser función de ***p(t)***

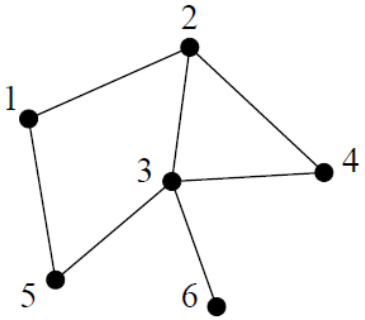
$$p_i(t+1) = \sum_{j=1}^{N} \frac{1}{k_j} a_{ij} \, p_j(t)$$

prob que esté en *nodo-j* a tiempo *t*

prob de elegir el enlace que lleva de *j* a *i*

Donde está Juan?

$$p_i(t+1) = \sum_{j=1}^{N} \frac{1}{k_j} a_{ij}\, p_j(t)$$



$$\boldsymbol{p}(t+1) = A\, D^{-1}\ \boldsymbol{p}(t)$$

matriz de adyacencia con columnas normalizadas por grado del nodo

$$A = \begin{matrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{matrix}$$

$$AD^{-1} = \begin{matrix} 0 & 1/3 & 0 & 0 & 1/2 & 0 \\ 1/2 & 0 & 1/4 & 1/2 & 0 & 0 \\ 0 & 1/3 & 0 & 1/2 & 1/2 & 1 \\ 0 & 1/3 & 1/4 & 0 & 0 & 0 \\ 1/2 & 0 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 1/4 & 0 & 0 & 0 \end{matrix}$$

Donde está Juan?

$$p_i(t+1) = \sum_{j=1}^{N} \frac{1}{k_j} a_{ij}\, p_j(t)$$

$$\boldsymbol{p}(t+1) = A\, D^{-1}\ \boldsymbol{p}(t)$$

Notar que si sucede que la distribución de probabilidad converge, es decir $\boldsymbol{p}(t \to \infty) \to \boldsymbol{p}_\infty$, se debe satisfacer que

$$\boldsymbol{p}_\infty = A\, D^{-1}\, \boldsymbol{p}_\infty$$

Esto significa que $\boldsymbol{p}_\infty$ es ***autovector*** de $A\, D^{-1}$, con ***autovalor*** 1.

Notas Algebra lineal:

- 1 es autovalor dominante de $A\, D^{-1}$, lo cual es razonable porque no esperamos que la probabilidad total deje de estar normalizada al evolucionar el tiempo.

Donde está Juan?

$$p_i(t+1) = \sum_{j=1}^{N} \frac{1}{k_j} a_{ij}\, p_j(t)$$

$$\boldsymbol{p}(t+1) = A\, D^{-1}\ \boldsymbol{p}(t)$$

Notar que si sucede que la distribución de probabilidad converge, es decir $\boldsymbol{p}(t \to \infty) \to \boldsymbol{p}_\infty$, se debe satisfacer que

$$\boldsymbol{p}_\infty = A\, D^{-1}\, \boldsymbol{p}_\infty$$

Esto significa que $\boldsymbol{p}_\infty$ es ***autovector*** de $A\, D^{-1}$, con ***autovalor*** 1.

Por inspección se puede ver que el vector de probabilidades de componente i-esima:

$$(p_\infty)_i = \frac{k_i}{\sum_{j=1}^{N} k_j}$$

es la solución buscada

a tiempos largos Juan está en…

$$\boldsymbol{p}_\infty = A\, D^{-1}\, \boldsymbol{p}_\infty \qquad\qquad (p_\infty)_i = \frac{k_i}{\sum_{j=1}^{N} k_j}$$

$$\frac{k_i}{\sum_{j=1}^{N} k_j} = \sum_{j=1}^{N} \frac{1}{k_j}\, a_{ij}\, \frac{k_j}{\sum_{r=1}^{N} k_r}$$

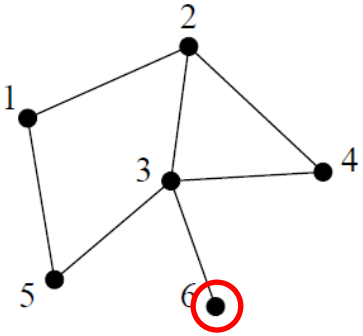
$$= \frac{1}{\sum_{r=1}^{N} k_r} \sum_{j=1}^{N} \frac{1}{k_j}\, a_{ij}\, k_j = \frac{1}{\sum_{r=1}^{N} k_r} \sum_{j=1}^{N} a_{ij}$$

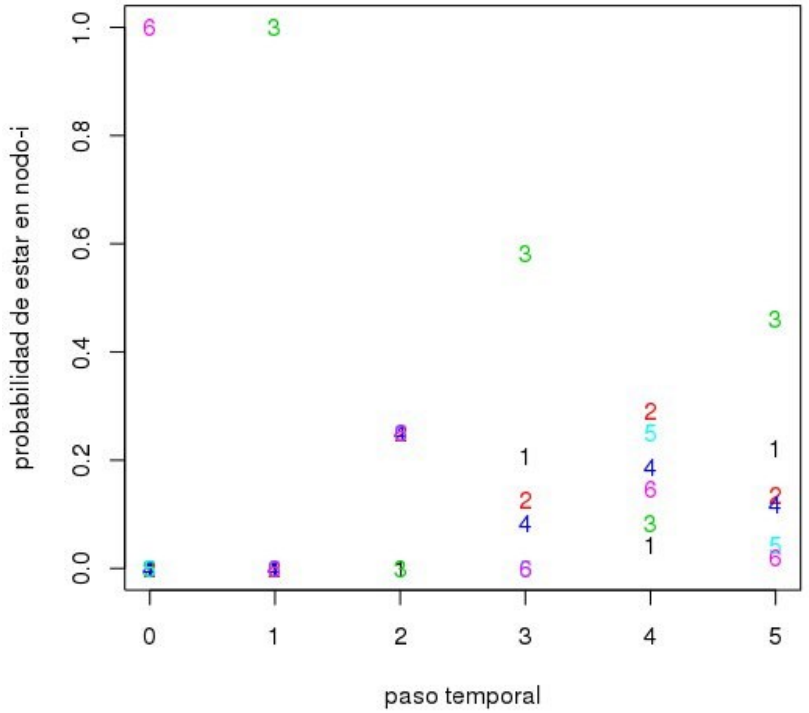
$$= \frac{k_i}{\sum_{j=1}^{N} k_j} \qquad \checkmark$$

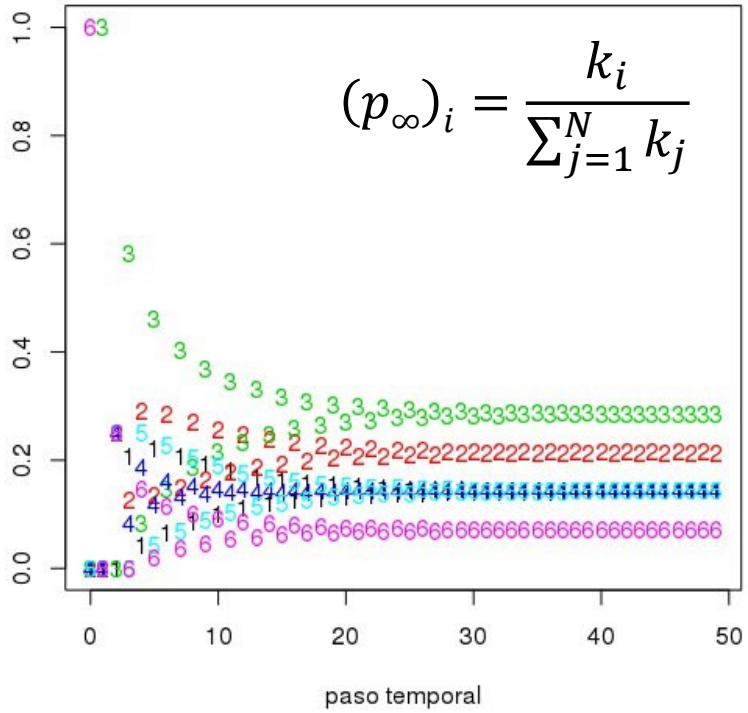
a tiempos largos Juan está en…

$$\boldsymbol{p}(0) = (0,0,0,0,0,1)^T$$

$$p_i(t+1) = \sum_{j=1}^{N} \frac{1}{k_j} a_{ij}\, p_j(t) \qquad \boldsymbol{p}(t+1) = A\, D^{-1}\ \boldsymbol{p}(t)$$

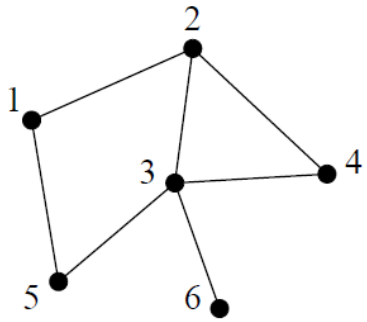






a tiempos largos Juan está en…

$$(p_\infty)_i = \frac{1}{\sum_{j=1}^{N} k_j} k_i$$



- La probabilidad de encontrar a Juan en el nodo-i es **proporcional al grado del nodo**

- En este proceso difusivo, o de caminata, las cosas tienden a **uniformizarse a nivel de enlaces**: En el régimen asintótico, la probabilidad de encontrar a Juan **atravesando uno de los M enlaces** de la red, $p_{enlace}(e_r=(i,j))$, es **uniforme**

$$(p_{enlace})_{ij} = \frac{k_j}{\sum_{r=1}^{N} k_r} \frac{1}{k_j} = \frac{1}{\sum_{r=1}^{N} k_r} = cte$$

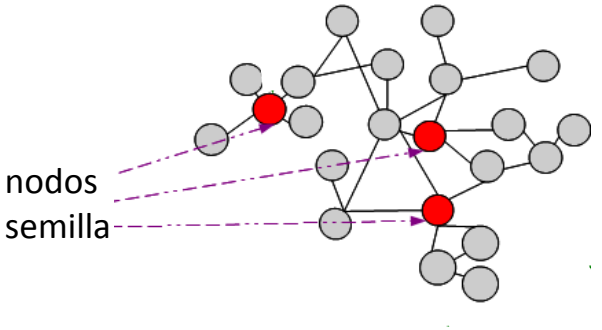
prob de estar en *j*

prob de elegir uno de los vecinos de *j*

Priorización basada en redes

- La probabilidad **asintótica** de encontrar a un caminante aleatorio en el nodo-i es **proporcional al grado del nodo**

$$(p_\infty)_i = \frac{1}{\sum_{j=1}^{N} k_j} k_i$$



- Con qué probabilidad, en una **caminata corta**, voy a visitar un nodo-x si a tiempo t=0 parto de un **conjunto nodos de interés (nodos semilla)?**

- Con qué probabilidad **asintótica** voy a visitar un nodo-x si a tiempo t=0 parto de un nodo tomado al azar de un conjunto de nodos de interés (nodos semilla) y eventualmente **fuerzo revisitar semillas**?

→ Pobabilidad de visita al nodo-x es una **medida de asociación** entre el nodo-x y el conjunto semilla

Integro dos espacios de conocimiento

1. el embebido en el conexionado de la red
2. el que utilicé para definir el conjunto de nodos-semilla

Priorización de genes asociados a enfermedades

ARTICLE

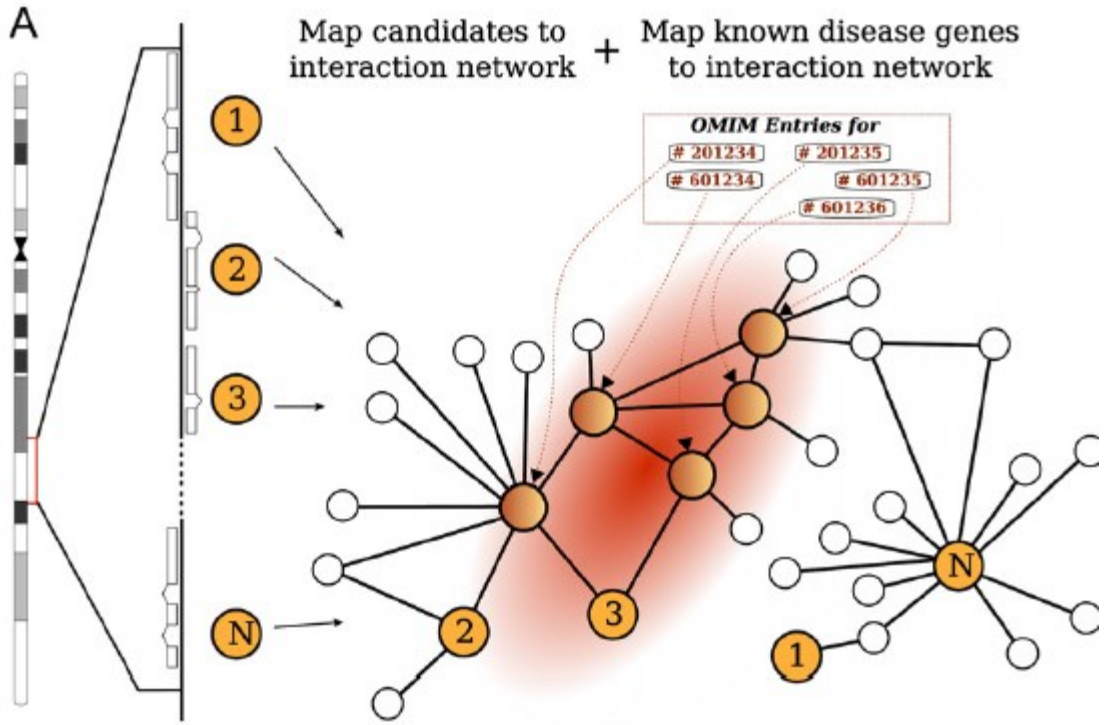
Walking the Interactome for Prioritization of Candidate Disease Genes

Sebastian Köhler,[1,2] Sebastian Bauer,[1,2] Denise Horn,[1] and Peter N. Robinson[1,*]

The identification of genes associated with hereditary disorders has contributed to improving medical care and to a better understanding of gene functions, interactions, and pathways. However, there are well over 1500 Mendelian disorders whose molecular basis remains unknown. At present, methods such as linkage analysis can identify the chromosomal region in which unknown disease genes are located, but the regions could contain up to hundreds of candidate genes. In this work, we present a method for prioritization of candidate genes by use of a global network distance measure, random walk analysis, for definition of similarities in protein-protein interaction networks. We tested our method on 110 disease-gene families with a total of 783 genes and achieved an area under the ROC curve of up to 98% on simulated linkage intervals of 100 genes surrounding the disease gene, significantly outperforming previous methods based on local distance measures. Our results not only provide an improved tool for positional-cloning projects but also add weight to the assumption that phenotypically similar diseases are associated with disturbances of subnetworks within the larger protein interactome that extend beyond the disease proteins themselves.

Kolher et al The American Journal of Human Genetics 2008

Priorización de nuevas asociaciones gen/enfermedad

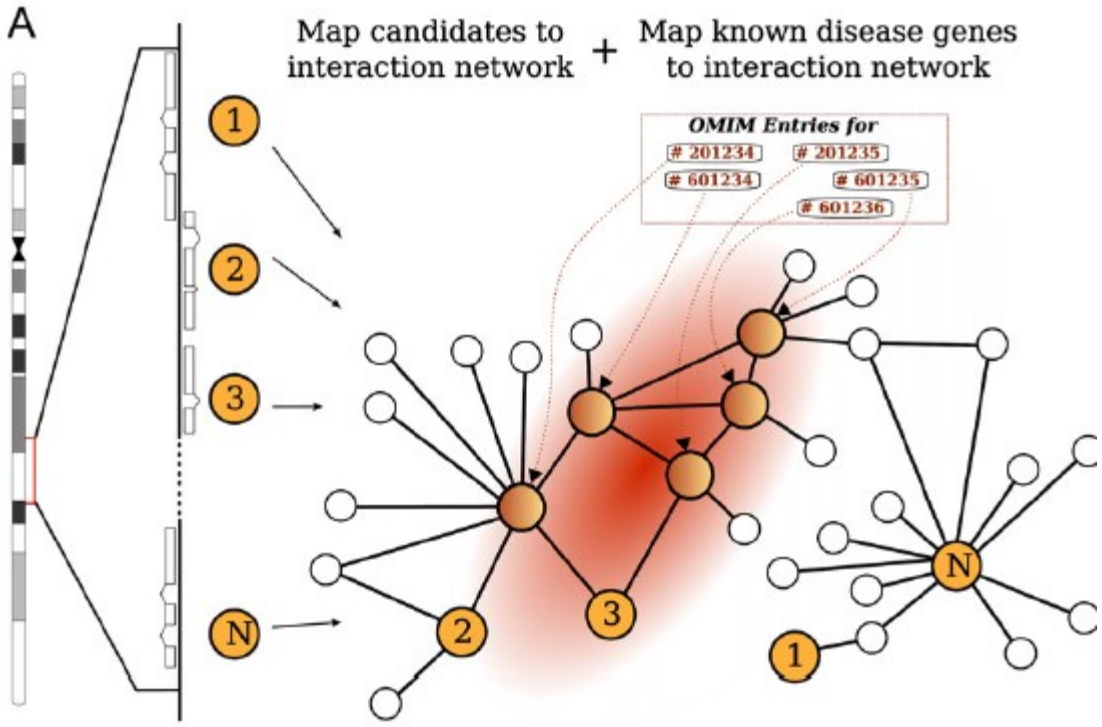


Algoritmos para **propagar sentido de pertenecia** al conjunto de interés:

Random Walk with Restart (Kholer 2009)
Net-Rank (Chen 2009)
Net-Propagation (Vanunu 2010)
Functional Flow (Navieba 2005)

Kolher et al The American Journal of Human Genetics 2008

Priorización de nuevas asociaciones gen/enfermedad



Algoritmos para **propagar sentido de pertenecia** al conjunto de interés:

Random Walk with Restart (Kholer 2009)

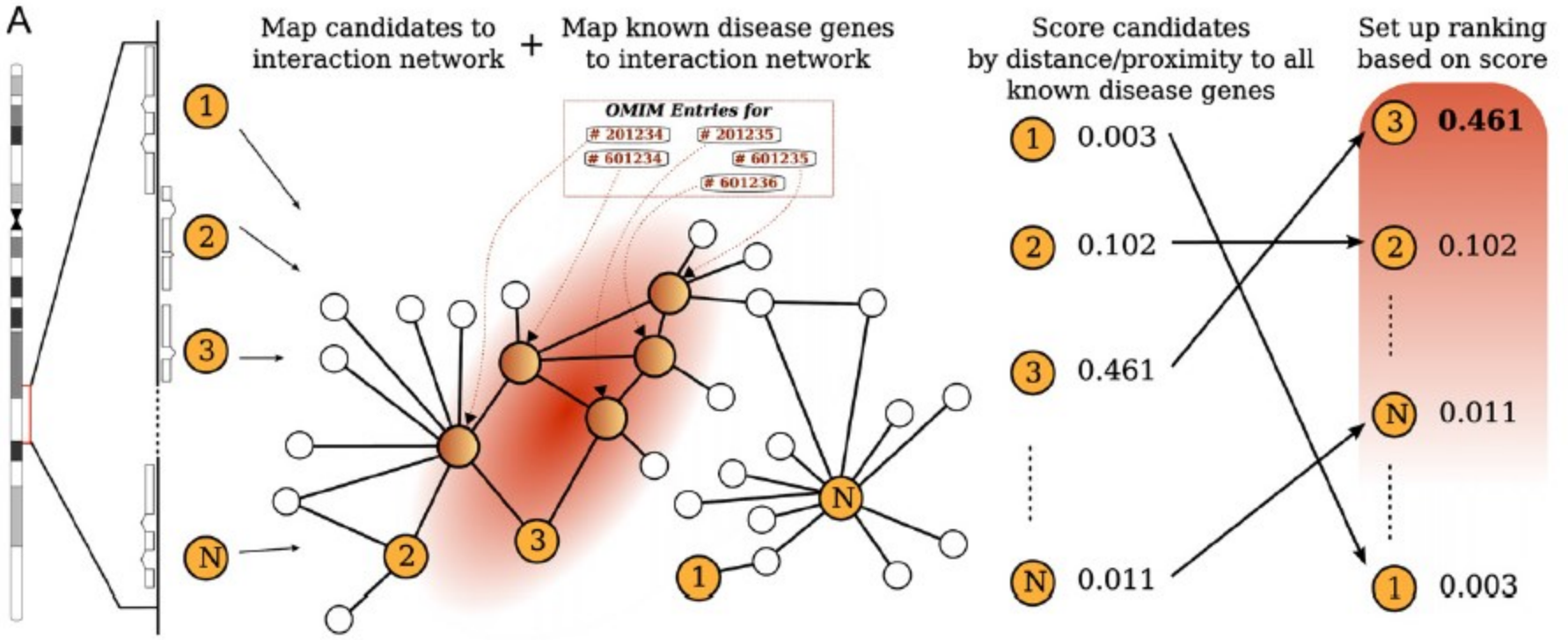
$$p_i^{t+1} = (1-r)\sum_{j=1}^{N} a_{ij} p_j^t + r\, p_i^{t=0}$$

$$\boldsymbol{p}^{t+1} = (1-\mathrm{r})\,\mathbf{A}\,\boldsymbol{p}^t + r\,\boldsymbol{p}^0$$

- r controla el sesgo hacia las condiciones iniciales
- el score se suele estimar a partir de simular un número reducido de pasos k: $\boldsymbol{p}^{\infty} \sim \boldsymbol{p}^{k}$

Kolher et al The American Journal of Human Genetics 2008

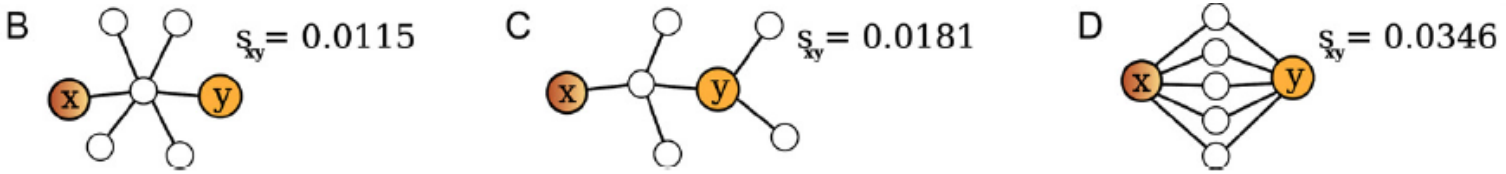
Priorización de nuevas asociaciones gen/enfermedad



$$\boldsymbol{p}^{t+1} = (1 - \mathrm{r})\,\mathbf{A}\,\boldsymbol{p}^{t} + r\,\boldsymbol{p}^{0}$$

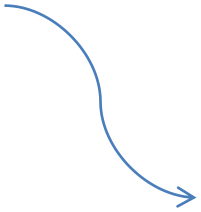
caminata aleatoria con *reset.*

El score se suele estimar a partir de simular k-pasos



Genómica funcional

- En la era post-genómica conocemos en alto grado la composición exacta del genoma de un gran número de organismos
- Sin embargo, entender qué función cumple la proteína asociada a un dado gen sigue siendo un reto
- En general los métodos computacionales utilizaban exhaustivamente el criterio de **similaridad de secuencia** para asignar funciones putativas a una dada proteina aun no caracterizada funcionalmente
- La disponibilidad de datos omicos a escala global permite abordar esto desde otra perspectiva:



BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i302–i310
doi:10.1093/bioinformatics/bti1054

Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps

Elena Nabieva[1,2], Kam Jim[2], Amit Agarwal[1], Bernard Chazelle[1] and Mona Singh[1,2,]*

[1]Computer Science Department and [2]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA

Received on January 15, 2005; accepted on March 27, 2005

Functional Flow

BIOINFORMATICS
Vol. 21 Suppl. 1 2005, pages i302–i310
doi:10.1093/bioinformatics/bti1054

Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps

Elena Nabieva[1,2], Kam Jim[2], Amit Agarwal[1], Bernard Chazelle[1] and Mona Singh[1,2,*]

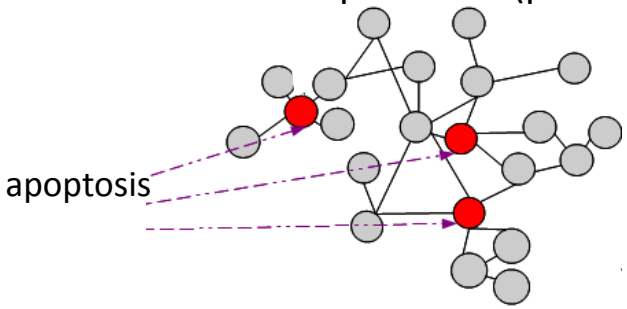
[1]Computer Science Department and [2]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA

Received on January 15, 2005; accepted on March 27, 2005

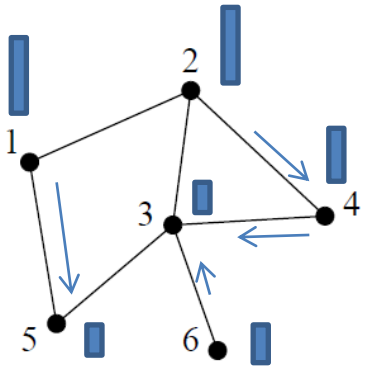
Asignación de funcionalidad a proteinas aun no caracterizadas simulando flujos de información en redes de interaccion de proteinas.

Dime con quien andas y te diré que haces

Red de interacción de proteinas (pesada)



1. Consideramos una función/proceso biológico (por ejemplo: GO-apoptosis)
2. Cada **proteina anotada** a esa categoría es considerada una **fuente de flujo** de una cantidad extensiva $x_i(t)$ que representa un *score-funcional*
3. Pensamos en un proceso difusivo en el que el flujo desde el *nodo-j* al *nodo-i* es **proporcional** a la diferencia de material: $flujo_{j->i} \sim C * (x_j - x_i)$
4. Luego de cada paso, se reinyecta ***x*** a las semillas.
5. La cantidad de flujo que recibe una proteina luego de k pasos de difusión representa el grado de asociación de la misma con la categoría funciopnal analizada.



$$\frac{dx_i}{dt} = C \sum_{j=1}^{N} w_{ij} \, (x_j - x_i) \, \theta((x_j - x_i))$$

Referencias

- *Networks, An Introduction*. Mark Newman
- Charlas
  - Peter Dodds – Random Walk and diffusion
  - Vito Latora

Random Walk with Restart (Kholer 2009)
Net-Rank (Chen 2009)
Net-Propagation (Vanunu 2010)
Functional Flow (Navieba 2005)

Prioritization (Lu 2012 – Physics Report)