

Redes y Sistemas Complejos
Cuarto Curso del Grado en Ingeniería Informática
Tema 4: Algoritmos de Poda y Visualización de Redes

Oscar Cordon García

Dpto. Ciencias de la Computación e Inteligencia Artificial
ocordon@decsai.ugr.es

INTRODUCCIÓN

Necesidad de la simplificación y visualización de redes (1)

La **sobrecarga de información** (*information overload*) se ha convertido en un problema fundamental debido al crecimiento exponencial de la información accesible en la sociedad actual

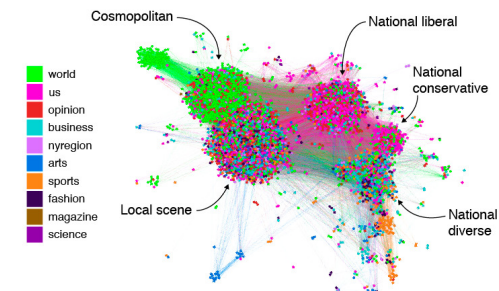
Es obligatorio considerar métodos de filtrado y compartición de información para resolver este problema



La **visualización de información** tiene el potencial necesario para ayudar a las personas a acceder a la información necesaria de una forma más eficaz e intuitiva

Definición R.A.E: “Representar mediante imágenes ópticas fenómenos de otro carácter. || Formar en la mente una imagen visual de un concepto abstracto”

La visualización de información comprende dos aspectos directamente relacionados: **Modelado estructural** y **Representación gráfica**



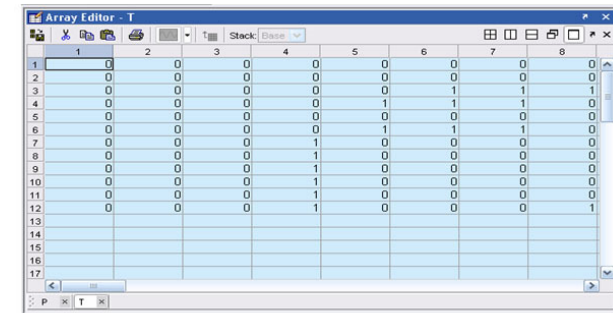
C. Chen. Information Visualization. 2nd edition. Springer 2004

INTRODUCCIÓN

Necesidad de la simplificación y visualización de redes (2)

El objetivo del **modelado estructural** es detectar, extraer y simplificar las relaciones subyacentes en nuestro dominio de aplicación

Estas relaciones forman una estructura que caracteriza el sistema o conjunto de datos estudiado. Ej: una pregunta típica es **¿cuál es la estructura de una red?**

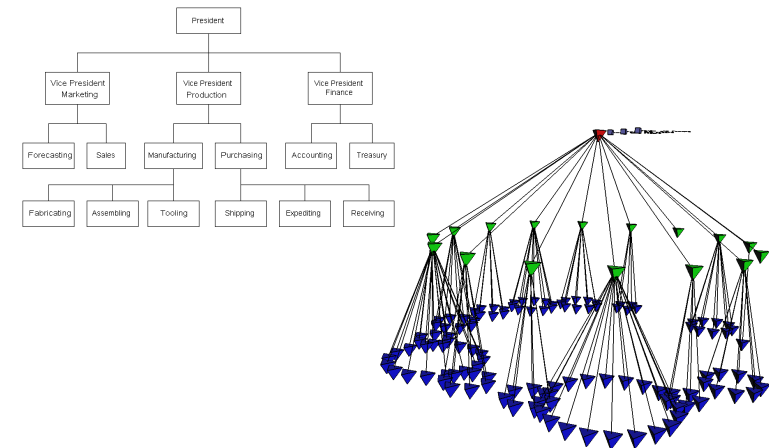


	1	2	3	4	5	6	7	8
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	1	1	1
4	0	0	0	0	1	0	1	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	1	1	1	0
7	0	0	0	1	0	0	0	0
8	0	0	0	1	0	0	0	0
9	0	0	0	1	0	0	0	0
10	0	0	0	1	0	0	0	0
11	0	0	0	1	0	0	0	0
12	0	0	0	1	0	0	0	1
13								
14								
15								
16								
17								

Fig. 3. Matriz T valores esperados.

Por el contrario, el propósito de la **representación gráfica** es transformar un modelo previo de una estructura en un modelo gráfico que permita examinar visualmente la estructura original e interactuar con ella

Ej: una estructura jerárquica se puede representar como un árbol cónico o un grafo hiperbólico

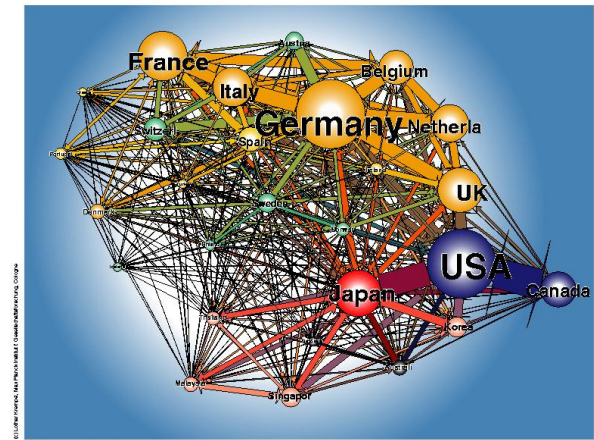


Las dos etapas suelen estar bastante mezcladas

En el dominio de la **visualización de redes complejas**, la gran dimensión de las redes genera dificultades para obtener representaciones gráficas útiles para el análisis

Problemática:

1. **Calidad:** cuanto más grande es la red, más probabilidad hay de que existan **errores en los datos**
2. **Complejidad:** **más variables**, más detalle, **más categorías**
3. **Velocidad:** Hoy en día, el interés se centra en obtener resultados de nuestra red lo bastante rápido como para que pueda considerarse un proceso **interactivo**
4. **Análisis:** ¿Qué **orden de complejidad** se requiere para los algoritmos que manejan redes de gran tamaño?



INTRODUCCIÓN

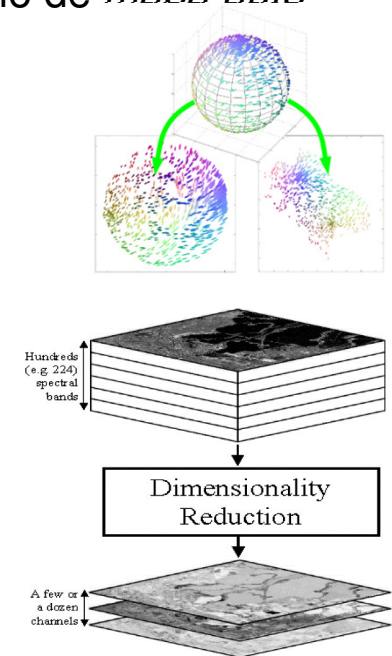
Necesidad de la simplificación y visualización de redes (4)

La solución se basa procesar las matrices de adyacencia de la red (cuyas entradas indican **similitud** o **distancia** entre las componentes) usando **técnicas de reducción de la dimensión** que faciliten la visualización de la red. Este tipo de técnicas se pueden clasificar en dos grupos:

1. Técnicas de naturaleza estadística, basadas en el análisis multivariante, como el *clustering*, el *análisis de componentes principales* (PCA) y el *escalado multidimensional* (MDS)
2. Técnicas de naturaleza conexionista, basadas habitualmente en el modelo de *mapa auto-organizativo* (SOM) de red neuronal

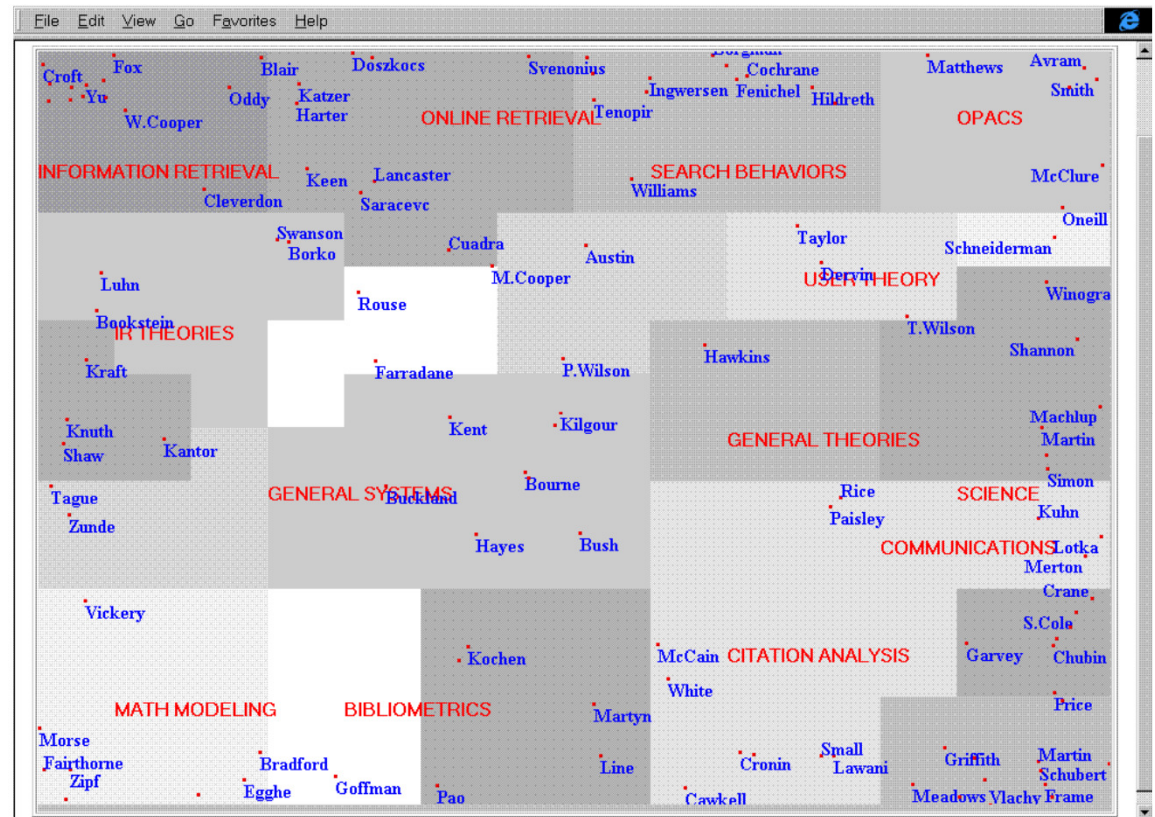
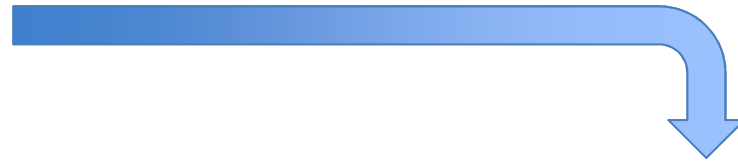
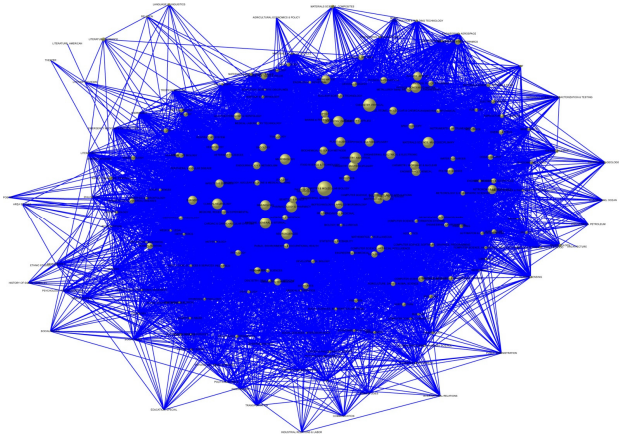
Su función es simplificar el complejo patrón de asociación existente entre las entidades de la red proyectando su gran número de dimensiones en un número mucho más pequeño, normalmente dos o tres, las procesables gráficamente por el ser humano (representaciones 2D y 3D)

Esa reducción suele implicar una pérdida de información como consecuencia del proceso de agregación de las múltiples dimensiones. Además, estas técnicas **no muestran los enlaces y sólo representan las relaciones individuales (locales) mediante proximidad espacial**



INTRODUCCIÓN

Necesidad de la simplificación y visualización de redes (4)



Ej: redes de cocitación científica
(White, Lin y McCain. 1998)

Alternativamente, existen **métodos de poda de redes** basados en la **eliminación de nodos y/o enlaces** que sacrifican la pérdida de información para obtener una ganancia en simplicidad

Se pueden hacer distintos tipos de poda. La más básica en una red ponderada es una **poda por umbral**, eliminando directamente aquellos nodos/enlaces cuyo valor/peso sea inferior o superior al valor límite. Sin embargo, puede provocar la pérdida de la conectividad de la red

También se puede reducir el número de enlaces de la red usando su **árbol generador minimal** (*minimum spanning tree*, **MST**). Esta filosofía realiza una poda más radical (cada par de nodos queda conectado por un único enlace) pero desde una perspectiva más global

http://decsai.ugr.es/~ta_ii/algoritmos/seccion.php?id=teoria&subseccion=applets/kef5_4

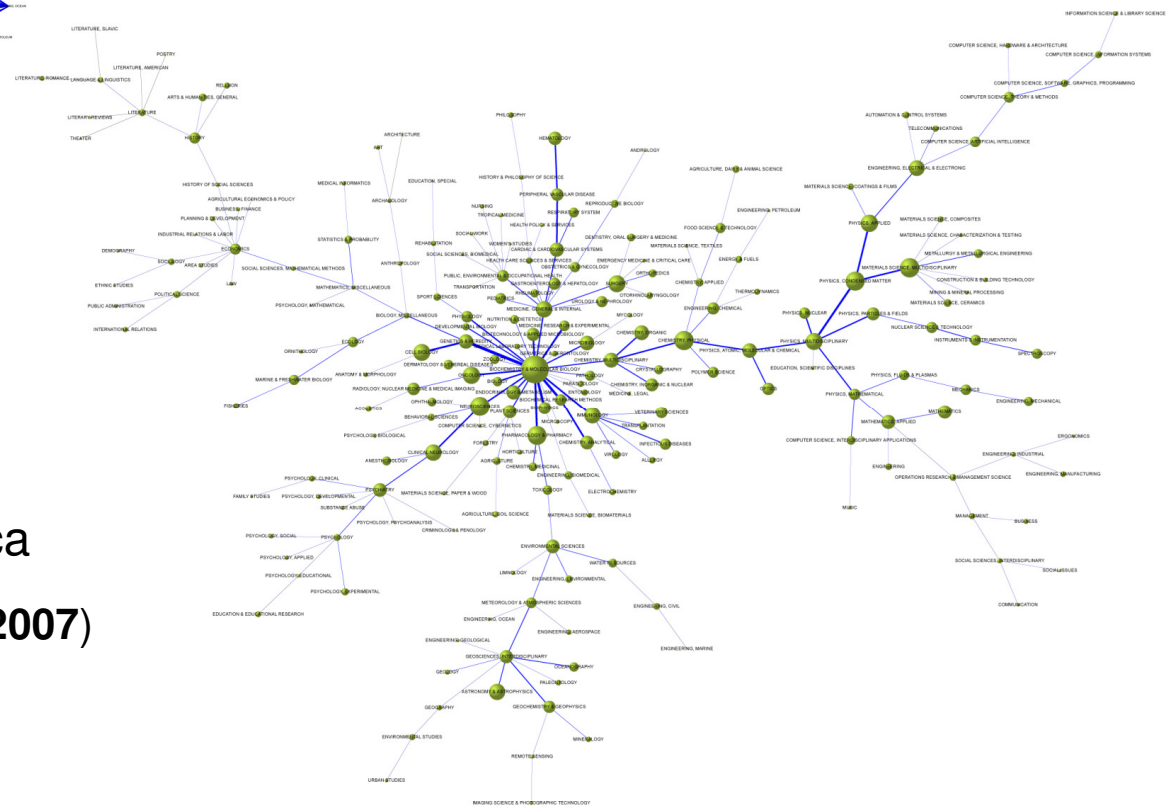
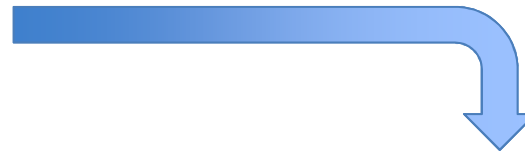
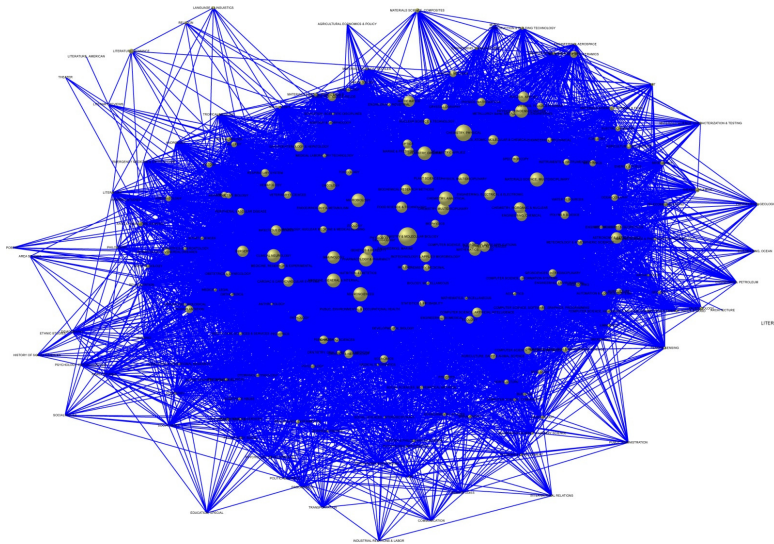
Finalmente, otra alternativa reduce la red original a una de sus **redes Pathfinder (PFNETs)**. El algoritmo Pathfinder se desarrolló en el marco de la ciencia cognitiva. Su función es construir varias redes distintas formadas sólo por los enlaces más relevantes de la red original según se satisfaga la desigualdad triangular en caminos de un tamaño determinado

Schvaneveldt, R.W., et al. 1989. Network structures in proximity data. En G. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory, Vol. 24, pp. 249–284. Academic Press

http://en.wikipedia.org/wiki/Pathfinder_network

INTRODUCCIÓN

Necesidad de la simplificación y visualización de redes (6)



Ej: redes de cocitación científica
(Vargas-Quesada y Moya-Anegón, 2007)

INTRODUCCIÓN

Necesidad de la simplificación y visualización de redes (7)

Una misma red puede representarse gráficamente de diversas formas → uso de distintos algoritmos de distribución (*layout*), con distintas filosofías

De dicha representación dependerá la facilidad con la que las personas dedicadas a analizarla puedan comprenderla

Dos métodos muy extendidos, basados en la filosofía *spring embedders* dentro de los *métodos de layout dirigidos por fuerzas* son los de Kamada-Kawai y Fruchterman-Reingold

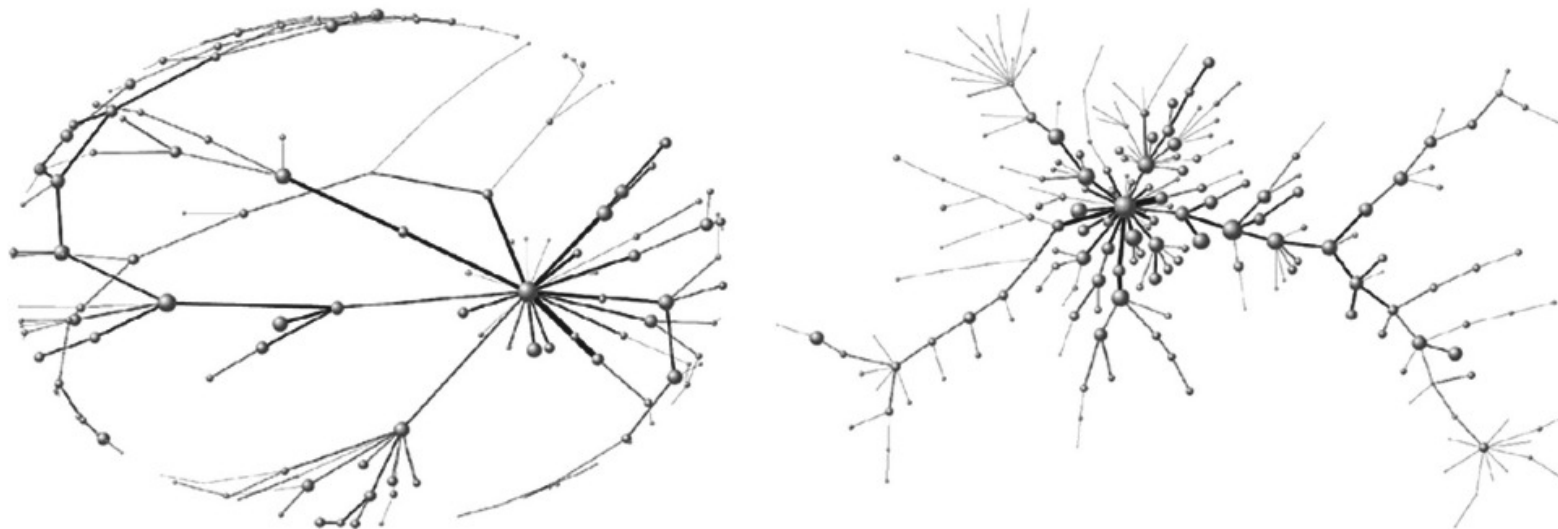
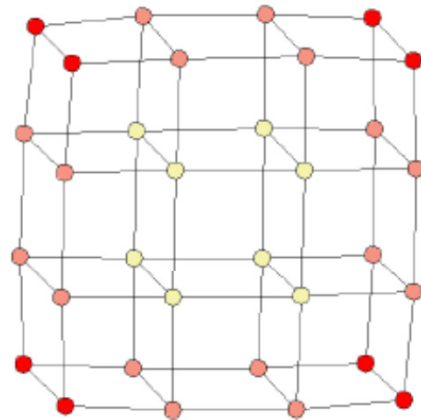
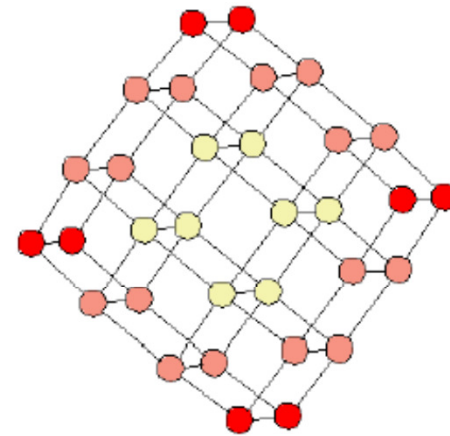


FIG. 1. Scientograms obtained using the algorithms of Fruchterman and Reingold (1991), and Kamada and Kawai (1989), respectively.

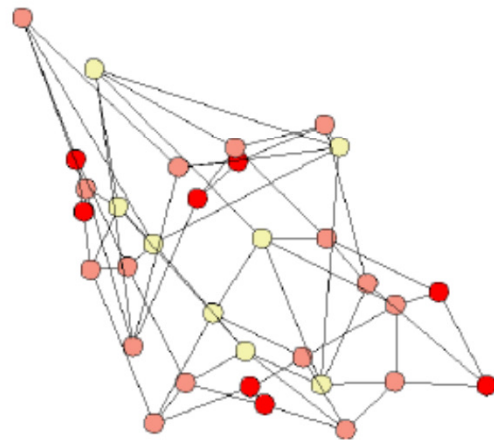
Cuboid (32 nodes & 64 edges)



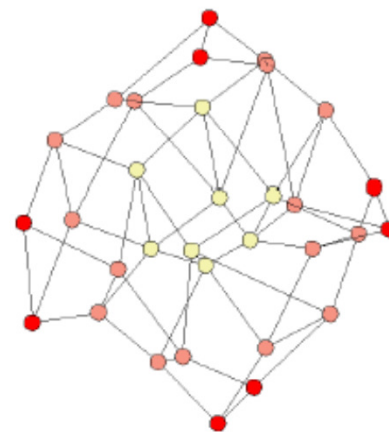
ODL



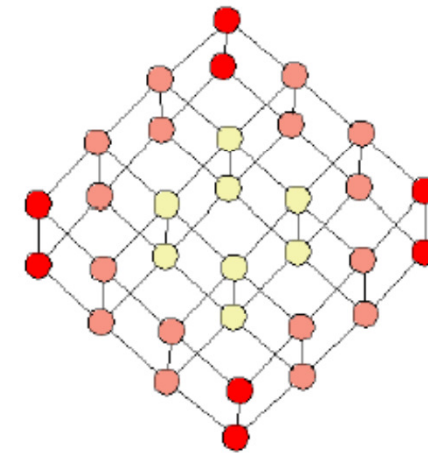
Frucherman-Reingold



Walshaw



ISOM

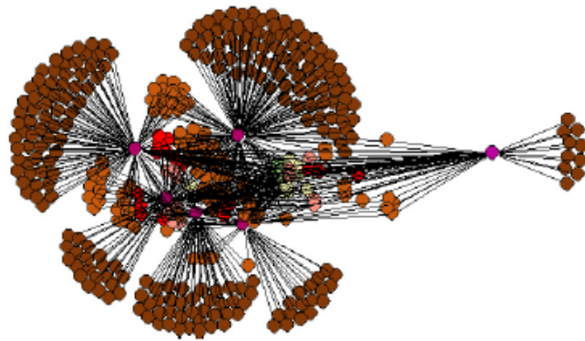


Kamada-Kawai

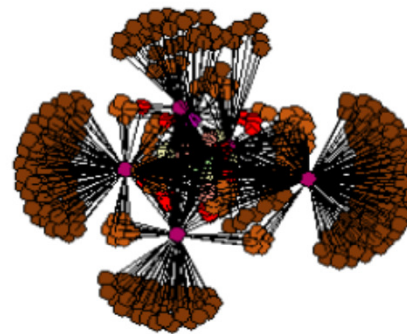
INTRODUCCIÓN

Necesidad de la simplificación y visualización de redes (9)

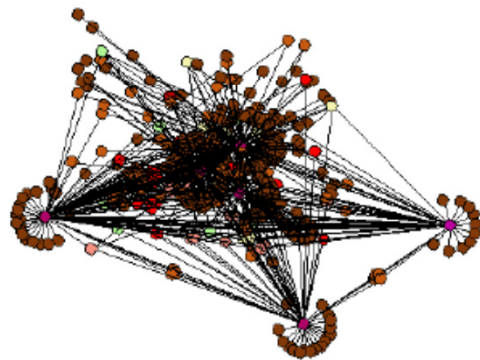
BGP AS map (369 nodes & 617 edges)



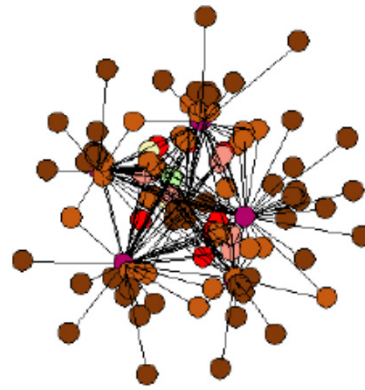
ODL



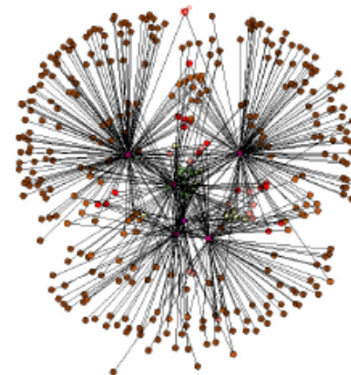
Fruchterman-Reingold



Walshaw



ISOM



Kamada-Kawai

In order to evaluate our hybrid layout algorithm on large networks, we focused on the problem of visualising Internet routing topologies. The role of routing protocols in networks is to ensure that information can be sent between computers connected to the network. Routing protocols are run internally to each of these networks (*intra-domain* routing) as well as between a network and its neighbour (*inter-domain* routing). Inter-domain routing in the Internet is coordinated by the Border Gateway Protocol (BGP) [11]. We have evaluated our layout algorithm using routing topologies generated by BGP as test data.

REDUCCIÓN DE LA DIMENSIÓN EN REDES

TÉCNICAS ESTADÍSTICAS DE REDUCCIÓN DE LA DIMENSIÓN EN REDES (1)

Técnicas de reducción de la dimensión basadas en el análisis multivariante:

1. *clustering*,
2. *análisis de componentes principales (PCA)* y
3. *escalado multidimensional (MDS)*

Los tres trabajan con una matriz de datos simétrica y cuadrada. Las filas y columnas de la matriz hacen referencia a las mismas entidades, y las celdas corresponden a un valor numérico que representa un **grado de asociación (similitud) o disociación (distancia)** entre las entidades

En el caso de clustering y MDS los valores de la matriz de entrada son “brutos”. En PCA se consideran covarianzas o correlaciones entre las dos variables

MDS y PCA proporcionan salidas parecidas, una nube de puntos (objetos) distribuidos en el espacio reducido (2D o 3D). El clustering devuelve generalmente un dendrograma bidimensional donde se representa una estructura jerárquica de clasificación

Los tres métodos pierden las relaciones locales al no representar los enlaces de la red

CLUSTERING JERÁRQUICO (1)

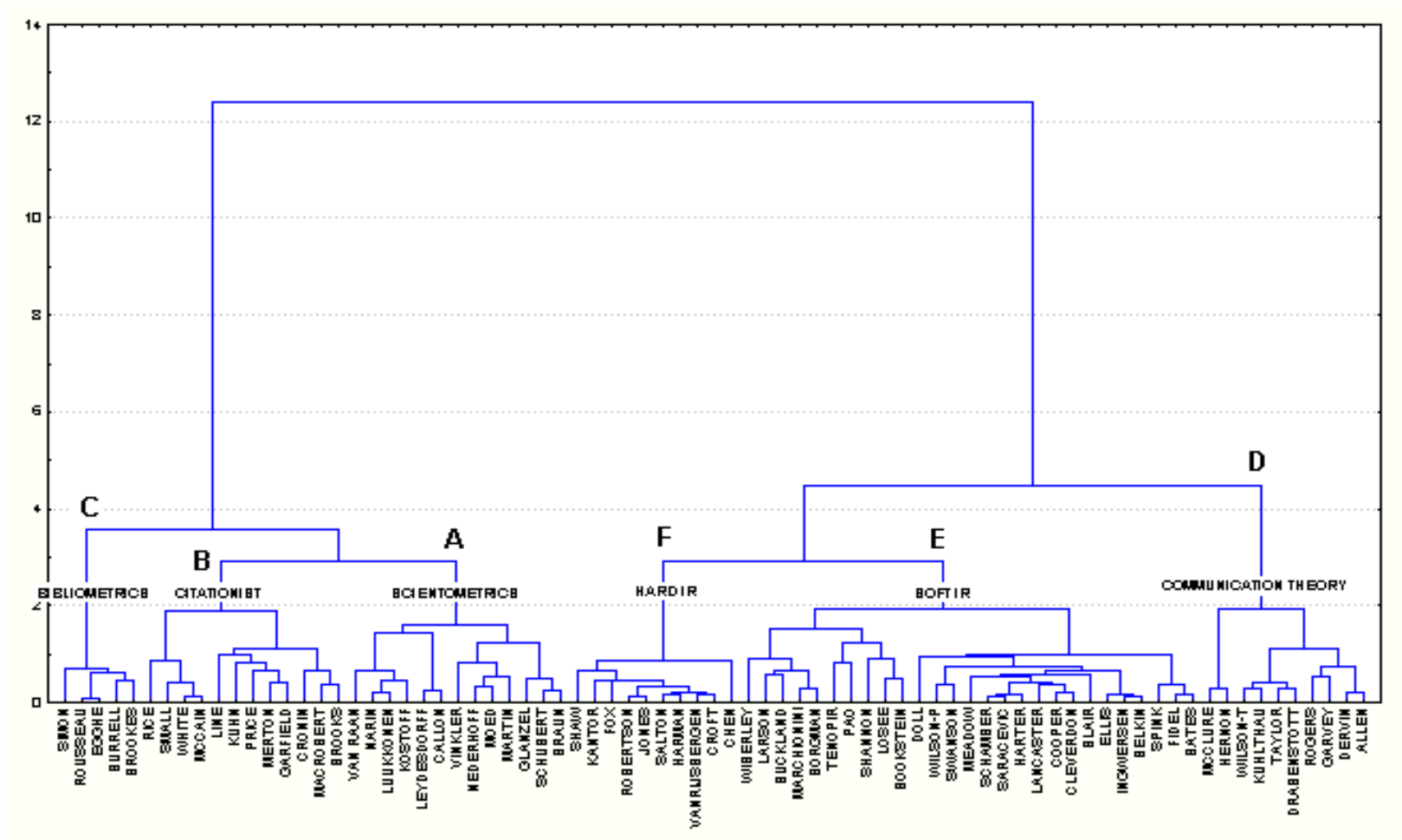
- El objetivo de los **métodos de clustering** es reducir el volumen de información mediante la categorización o agrupamiento de datos con características similares
- El clustering aporta herramientas que facilitan la construcción automática de taxonomías y minimizan la intervención humana en el proceso
- Esta técnica se utiliza para crear una gráfica bidimensional, denominada **dendrograma**, de agrupaciones (clusters) de diferentes objetos cuyas relaciones subyacen en la matriz de datos
- Existen alrededor de 150 técnicas diferentes de clustering, clasificadas en función del principio de aglomeración utilizado
- Cualquiera de esas técnicas de clustering utiliza al menos dos elementos: la **métrica de distancia** y las **reglas de aglomeración**
- La combinación de ambos factores permite clasificar los objetos partiendo de una configuración inicial en la que cada uno pertenece a una clase distinta independientemente de la distancia existente entre ellos

CLUSTERING JERÁRQUICO (2)

- A través de un proceso iterativo se van agrupando los objetos más próximos, al tiempo que se sube en el nivel de la jerarquía, lo que permite representar las agrupaciones en forma de árbol
- Entre las reglas de aglomeración más usadas encontramos:
 - agrupamiento simple (*single linkage*), también denominado **método del vecino más cercano** (*nearest neighbor method*),
 - agrupamiento completo (*complete linkage*) o **método del vecino más lejano** (*furthest neighbor method*),
 - **agrupamiento promedio** (*average linkage*), y
 - **método de Ward** o método de la suma de cuadrados (sum of squares method)
- **La regla de aglomeración óptima es aquella que permite la mejor interpretación de la estructura de cada espacio n-dimensional concreto**

CLUSTERING JERÁRQUICO (3)

Clustering basado en Ward de Co-citación de Autores en Biblioteconomía (Moya-Anegón et al. 1998)

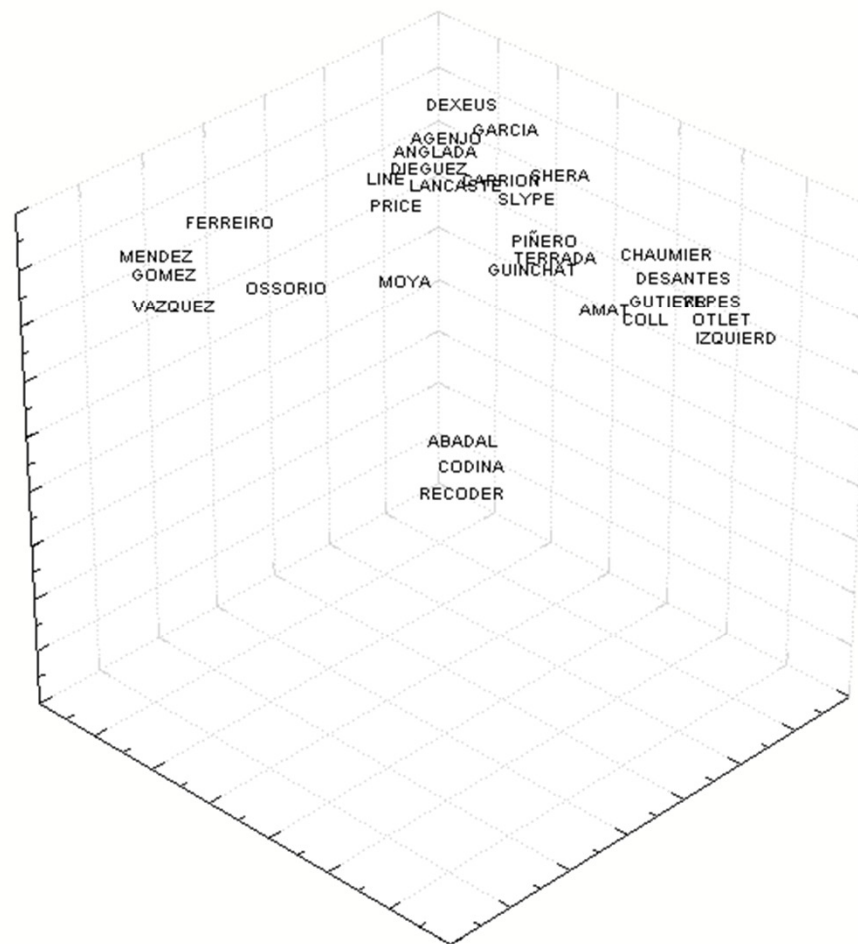


ANÁLISIS DE COMPONENTES PRINCIPALES (1)

- El **análisis factorial** es un método de análisis multivariante que intenta explicar un conjunto extenso de variables observables mediante un número reducido de variables hipotéticas llamadas **factores comunes** sin perder información ni capacidad explicativa
- El **PCA** se basa en que no hay factores comunes específicos sino que cada factor debe explicar el máximo de la variabilidad inicial de los datos
- Se asume que las variables son susceptibles de ser reducidas a factores comunes, es decir, que cada factor está presente en mayor o menor grado en todas las variables

ANÁLISIS DE COMPONENTES PRINCIPALES (2)

Mapa PCA 3D de Co-citación de Autores en Biblioteconomía (Moya-Anegón et al. 1998)



ESCALADO MULTIDIMENSIONAL (1)

- El **MDS** incluye una familia de métodos de escalado que establecen correspondencias entre datos de alta dimensión y espacios 2D/3D de forma que se mantienen las similitudes/distancias originales entre los datos
- La idea es **que las relaciones de distancia sigan siendo proporcionalmente consistentes** según se va reduciendo la dimensión del espacio
- Es posible capturar la naturaleza de un conjunto de datos a partir de los agrupamientos que emergen de la distribución espacial obtenida por el MDS
- Dado que su propósito es generar un mapa de objetos, puede considerarse como una alternativa al PCA. En cambio, suele combinarse con las técnicas de clustering

ESCALADO MULTIDIMENSIONAL (2)

- Los **métodos métricos de MDS** se basan en el análisis de valores propios de la matriz que muestran la relación entre cada elemento
- Si la distancia original entre dos nodos es d , la distancia en el espacio 2D/3D debe ser $\lambda \cdot d$, donde λ es un valor constante para todos los elementos de la red
- Proporcionan la solución exacta mediante el método de valores singulares pero no son iterativos y son bastante costosos
- Normalmente los datos se representan de acuerdo a los vectores propios correspondientes a los dos valores propios más grandes

ESCALADO MULTIDIMENSIONAL (3)

- Los **métodos no métricos de MDS** fueron propuestos por Kruskal para solucionar los problemas de los métodos métricos
- Consideran una medida estadística denominada **stress** para posicionar los nodos en el espacio de menor dimensión:

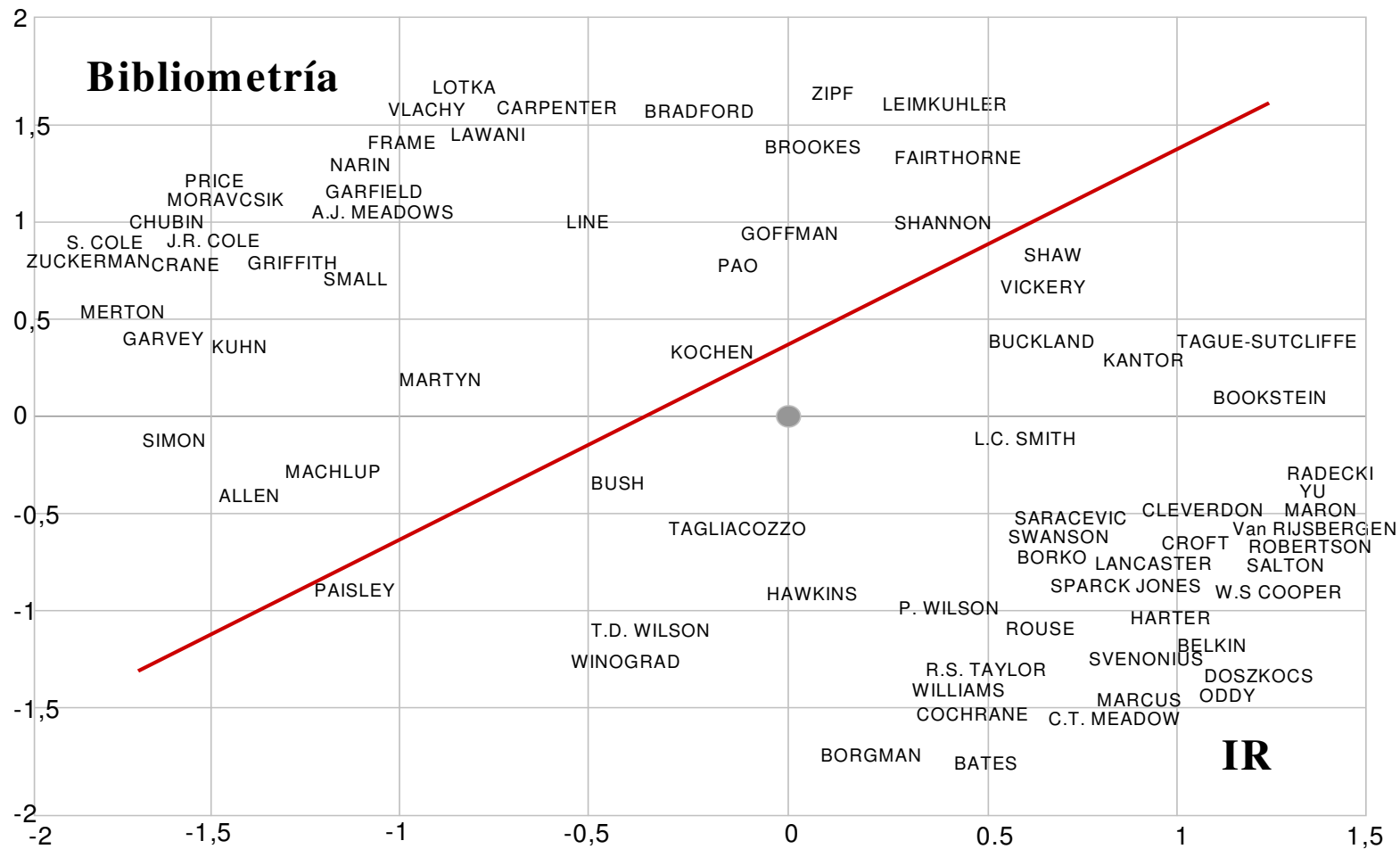
$$Stress = \frac{\sum_{i < j} (d_{ij} - g_{ij})^2}{\sum_{i < j} g_{ij}^2}$$

donde d_{ij} es la distancia en el espacio de alta dimensión y g_{ij} es la distancia en el espacio de baja dimensión

- La medida evalúa el grado de ajuste entre similitudes observadas y calculadas. Los nodos se van recolocando para reducir su valor. Se detiene el **proceso iterativo** cuando el valor global de *stress* queda por debajo de un umbral

ESCALADO MULTIDIMENSIONAL (4)

Mapa MDS 2D de Co-citación de Autores en Biblioteconomía (White y McCain 1998)



MAPAS AUTO-ORGANIZATIVOS (1)

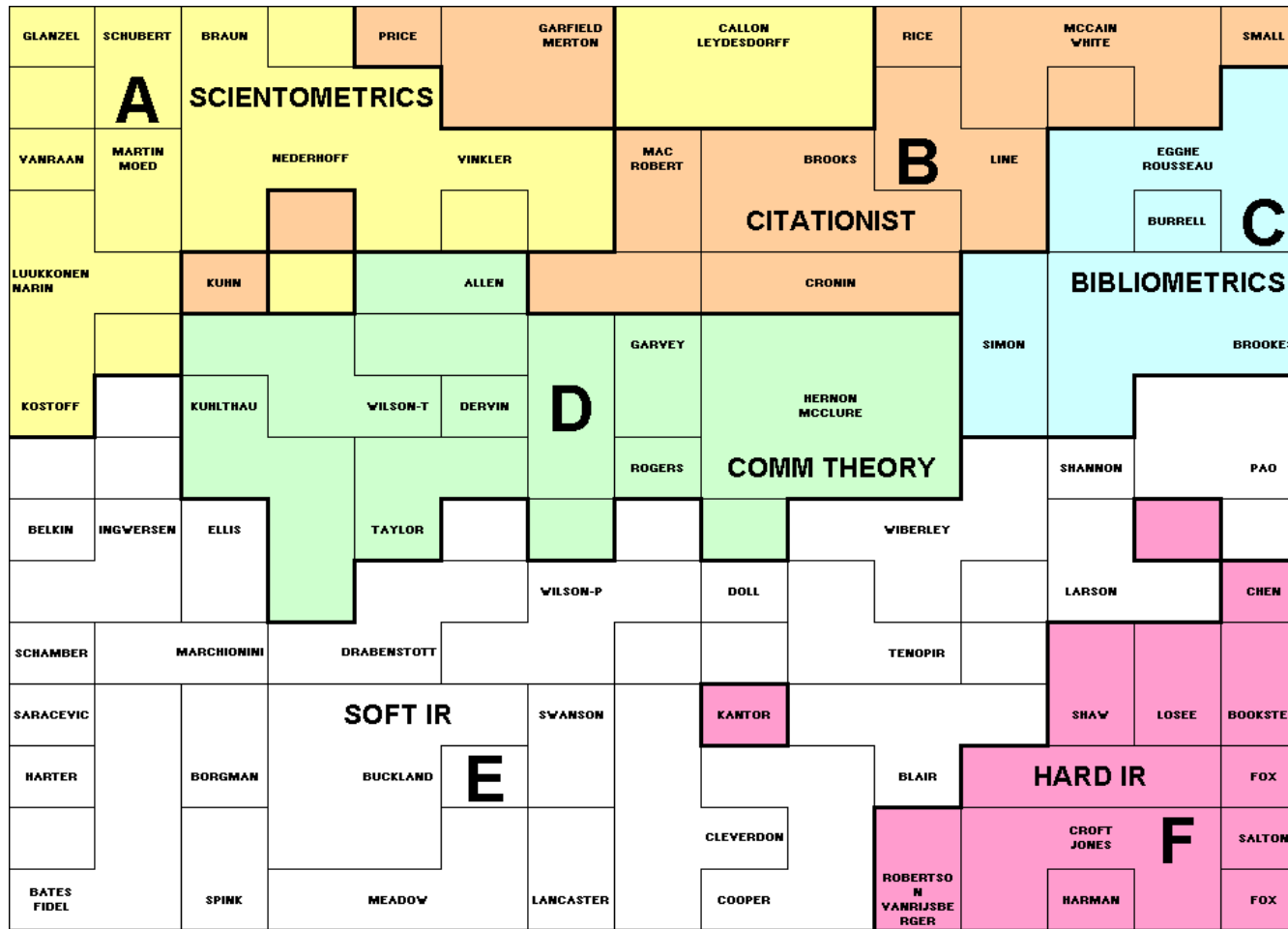
- En los 80, Teuvo Kohonen demostró que, suponiendo una estructura propia y una descripción funcional del comportamiento de una red neuronal **SOM**, una información de entrada por sí sola era suficiente para forzar la formación de mapas topológicos
- Su funcionamiento se basa en establecer una **correspondencia automática entre la información n-dimensional de entrada y un espacio 2D de salida**
- Los datos de entrada con características comunes activan zonas próximas del mapa, formado por neuronas de salida dispuestas en un plano
- Cuando se ingresa un dato a la red ésta reacciona de modo que solo una neurona de la capa de salida resulta activada, determinando un punto en el mapa
- Lo que realmente hace la red es clasificar la información de entrada ya que la neurona ganadora representa la clase a la que pertenece el dato. Ante entradas similares siempre se activa la misma neurona
- Por tanto, el SOM es válido para **establecer relaciones, desconocidas previamente, entre un conjunto determinado de datos**

MAPAS AUTO-ORGANIZATIVOS (2)

- El método de aprendizaje del modelo SOM se denomina competitivo y es de tipo no supervisado y *off-line*, por lo que presenta una etapa previa de entrenamiento y otra posterior de operación
- Posee dos limitaciones:
 1. el proceso de aprendizaje suele ser largo y arduo, y
 2. para aprender nuevos datos hay que repetir el proceso completo de aprendizaje
- No obstante, la versatilidad de este tipo de red es muy amplia y tiene varias ventajas:
 1. Visualización ordenada: ayuda a entender las estructuras subyacentes a los datos
 2. Visualización de clusters: se puede observar la densidad de clustering de las distintas regiones del mapa aunque no la forma y los límites de cada cluster
 3. Capacidad para manejar datos perdidos y para detectar *outliers*
- En un sentido estricto, **el SOM no es una técnica de reducción de la dimensión** como las anteriores. Sin embargo, la ordenación que realiza en el espacio bidimensional puede considerarse una verdadera dimensión reducida

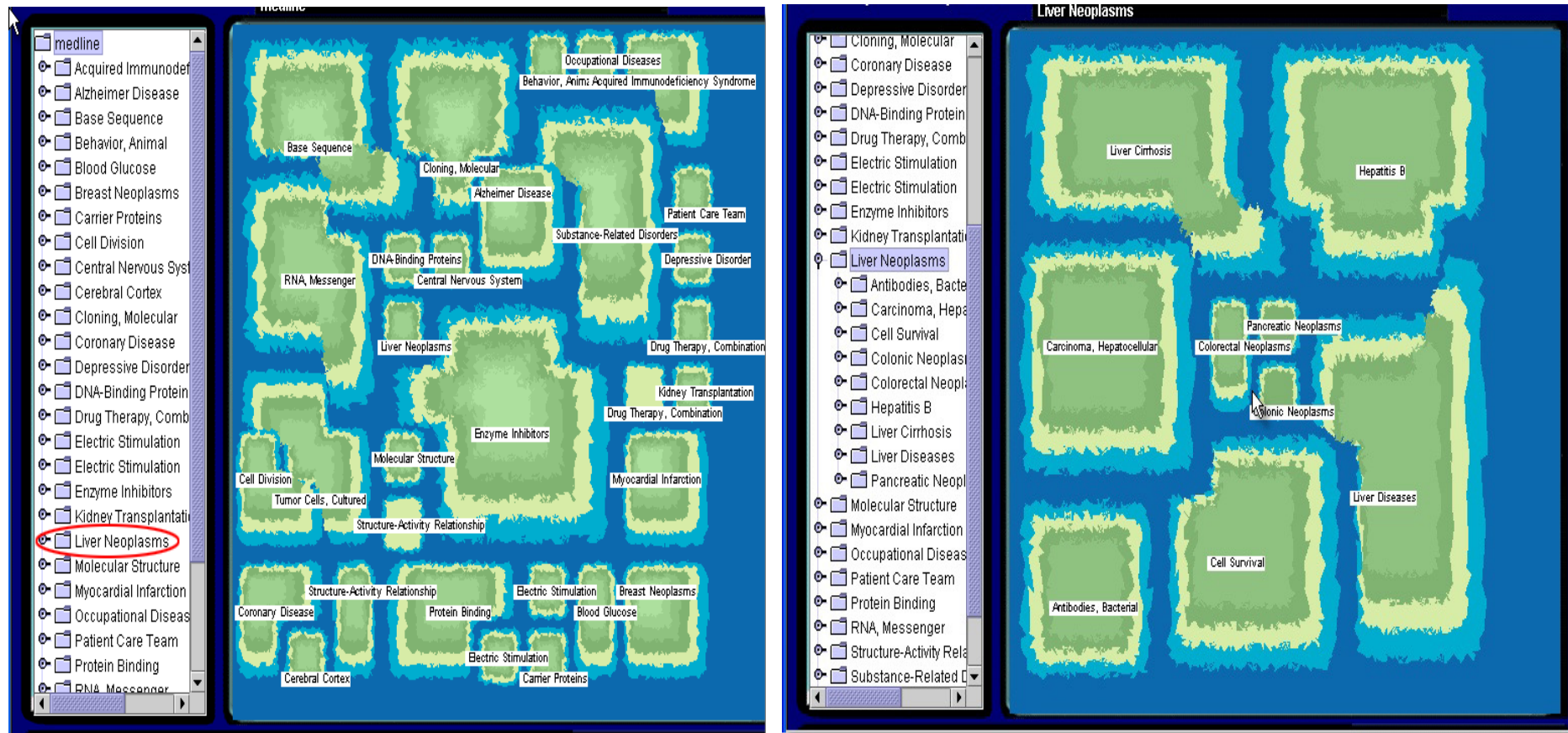
MAPAS AUTO-ORGANIZATIVOS (3)

Mapa SOM de Co-citación de Autores en Biblioteconomía (White y McCain 1998)



MAPAS AUTO-ORGANIZATIVOS (4)

Mapa SOM del Cáncer (Chen 1998)



PODA DE REDES

ESTRATEGIAS PARA LA VISUALIZACIÓN DE GRANDES REDES (1)

La gran dimensión de las redes genera dificultades para obtener representaciones gráficas útiles para el análisis. **Solución:** reducir el número de nodos y/o enlaces

1. Uso de umbrales:

- Nodos: sólo autores que hayan escrito al menos x artículos
- Nodos: sólo usuarios de Twitter con más de y seguidores (grado $> y$)
- Enlaces: sólo enlaces con peso $> z$

2. Uso de MSTs: visualizan todos los nodos con un subconjunto mínimo de enlaces

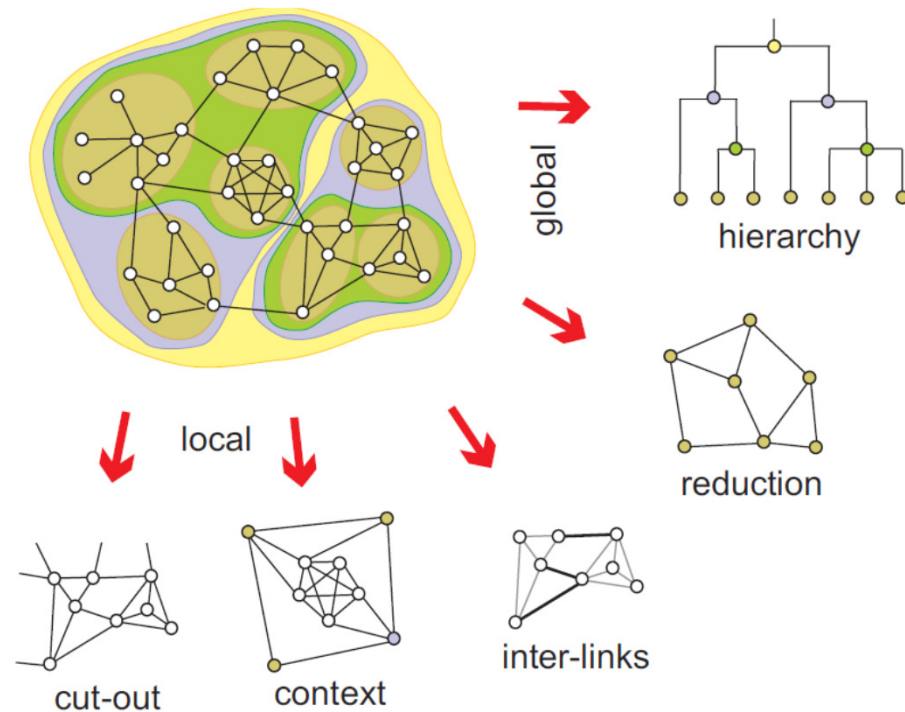
3. Uso de PFNETs: visualizan todos los nodos con los enlaces más significativos

- Uso de la desigualdad triangular para eliminar enlaces redundantes o contra-intuitivos
- Los enlaces restantes recogen mejor las relaciones semánticas de la red que los MSTs y representan de forma más precisa las relaciones locales que el MDS

4. Agregación de nodos: Se integran varios nodos en uno solo y se muestran únicamente los enlaces entre los clusters de nodos así creados

ESTRATEGIAS PARA LA VISUALIZACIÓN DE GRANDES REDES (2)

Enfoques para manejar redes complejas grandes (manual de Pajek):



With **Pajek** we can: *find* clusters (components, neighbourhoods of ‘important’ vertices, cores, etc.) in a network, *extract* vertices that belong to the same clusters and *show* them separately, possibly with the parts of the context (detailed local view), *shrink* vertices in clusters and show relations among clusters (global view).

ESTRATEGIAS PARA LA VISUALIZACIÓN DE GRANDES REDES (3)

Ejemplo: agregación de nodos en redes de interacciones entre proteínas (PPI)

B.-H. Ju, K. Han. Complexity management in visualizing protein interaction networks. *Bioinformatics* 2003, 19(s1): i177–i179

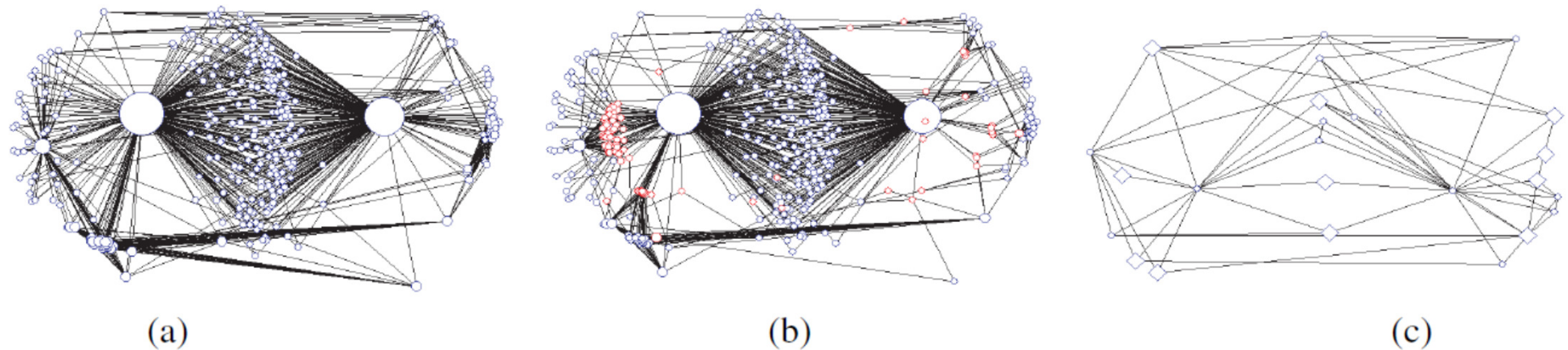


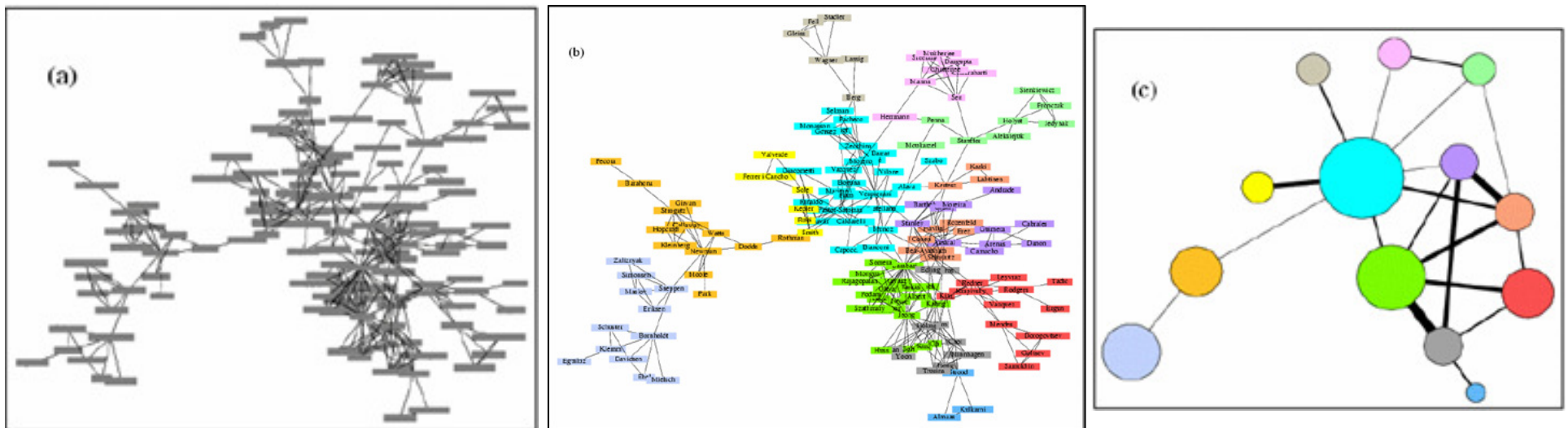
Fig. 1. (a) A protein interaction network with 307 nodes and 1063 edges. (b) Simplified graph with 332 nodes and 712 edges by replacing the cliques in Figure 1a with star-shaped subgraphs centered at dummy nodes, shown as red circles. (c) Simplified graph with 25 nodes and 62 edges by collapsing a group of nodes with the same interacting partners in Figure 1a into composite nodes, shown as diamonds.

ESTRATEGIAS PARA LA VISUALIZACIÓN DE GRANDES REDES (4)

Enfoque avanzado de agregación de nodos:

Red de co-autoría de Física. Aplicación del algoritmo de detección de comunidades de Newman y Girvan (2004). Agregación de todos los nodos de cada comunidad en un único nodo. Reducción del número de enlaces

El tamaño de cada nodo comunidad indica su número de autores. La anchura de los enlaces representa el número de enlaces de co-autoría entre las comunidades



El **método Pathfinder** es una técnica de modelado estructural desarrollada originalmente por Schvaneveldt, Durso y Dearholt (1989) para el análisis de datos de proximidad en Psicología

Schvaneveldt, R.W., et al. 1989. Network structures in proximity data. En G. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory, Vol. 24, pp. 249–284. Academic Press

La técnica se basa en construir una **red ponderada** a partir de estimaciones de proximidad entre pares de ítems y aplicar el algoritmo de poda para eliminar todos los enlaces de la red menos los más importantes desde una perspectiva global

Desde un punto de vista general, Pathfinder es un mecanismo de reducción de enlaces que preserva las relaciones semánticas más sobresalientes de la red. Se basa en los conceptos de **desigualdad triangular** y **camino mínimo en grafos**

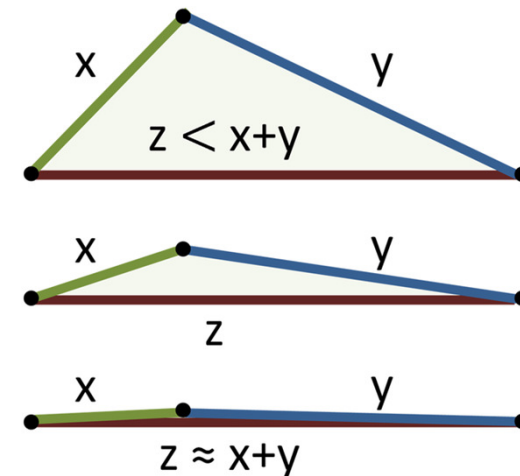
Las redes podadas resultantes se denominan **Pathfinder networks** (PFNETs). Pueden obtenerse varias PFNETs distintas a partir de la red inicial modificando los valores de los parámetros del algoritmo

Desigualdad triangular: En cualquier espacio, la distancia entre dos puntos x y z es como mucho tan grande como la suma de la distancia de x a y y la distancia de y a z :

$$d(x, z) \leq d(x, y) + d(y, z)$$

En el espacio Euclideo:

- se da la desigualdad estricta
- La distancia más corta entre dos puntos es siempre la línea recta

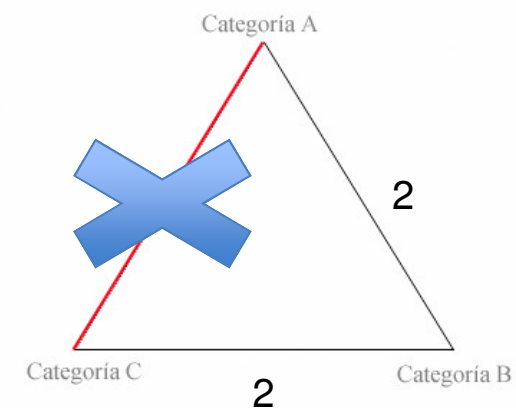


Idea fundamental del método Pathfinder:

“La eliminación de los enlaces que violan la desigualdad triangular en una red asegura la preservación de las distancias geodésicas entre los nodos y proporciona una estructura más simple que incluye precisamente los enlaces responsables de los caminos más económicos”

Un enlace que no verifica la desigualdad triangular no pertenecerá nunca a un camino geodésico. Se elimina al considerarlo **redundante**

La lógica es que si el significado de una relación semántica puede derivarse de forma más precisa o fiable a partir de otras relaciones, esa relación es redundante y puede eliminarse sin correr riesgos



Se elimina el enlace AC

La topología de una PFNET viene determinada por dos parámetros r y q :

- El peso (distancia) de un camino geodésico de la red P compuesto por k enlaces se calcula con la métrica de Minkowski, que depende de un parámetro $r \in [1, \infty]$:

$$w(P) = \left[\sum_{i=1}^k w_i^r \right]^{1/r}$$

$r=1 \rightarrow$ distancia camino = suma de los pesos de los arcos
 $r=2 \rightarrow$ distancia camino = distancia Euclidea
 $r=\infty \rightarrow$ distancia camino = **peso del enlace de mayor peso**

$$\lim_{r \rightarrow \infty} [w_i^r + w_j^r]^{1/r} = \text{maximum}(w_i, w_j)$$

- El parámetro $q \in \{2, \dots, n-1\}$ indica el número máximo de enlaces de un camino (sin ciclos) para el que se exige la satisfacción de la desigualdad triangular en la PFNET

A mayor valor de q , más exigente es la condición y más enlaces se podan

La red correspondiente se denomina $PFNET(r,q)$. Preserva la distancia geodésica y simplifica la estructura de la red original dependiendo del valor de q :

$$w_{n_1 n_k} \leq \left(\sum_{i=1}^{k-1} w_{n_i n_{i+1}}^r \right)^{1/r} \quad \forall k = 2, 3, \dots, q$$

Se pueden construir una serie de redes de complejidad decreciente aumentando el valor de $q \in \{2, \dots, n-1\}$

La $PFNET(r,q=1)$ correspondería a la red original. La $PFNET(r=1,q=n-1)$ está formada por el menor número posible de enlaces

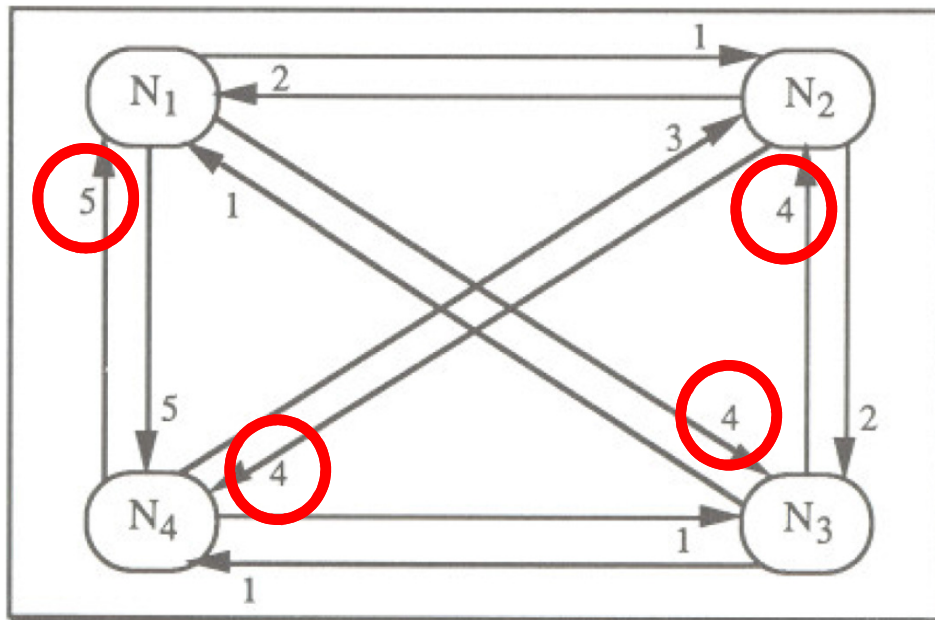
En dicha red, todos los caminos posibles son caminos geodésicos. Sin embargo, no corresponde al MST de la red original ya que puede haber redundancia de enlaces (varios enlaces con el mismo peso)

De hecho, la $PFNET(r=1,q=n-1)$ es la unión de todos los MSTs de la red original

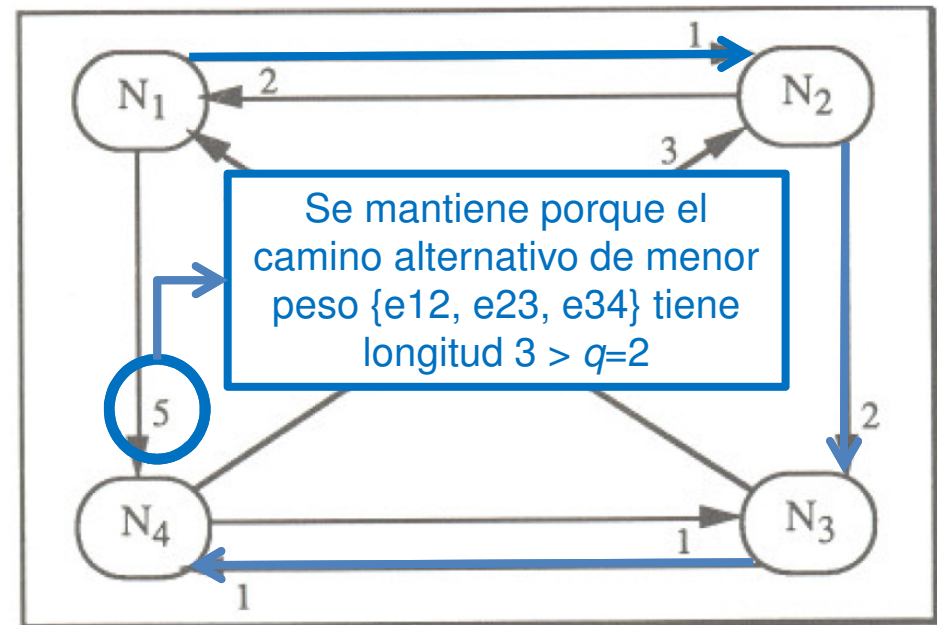
EL ALGORITMO PATHFINDER DE PODA DE REDES

Ejemplos (1)

Ejemplo de funcionamiento: PFNET($r=1, q=2$)



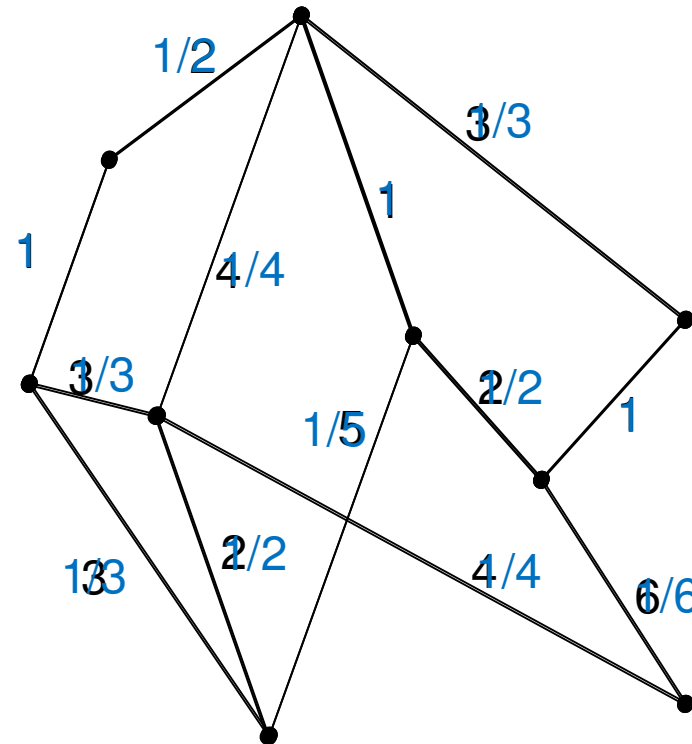
$$W = \begin{matrix} 0 & 1 & 4 & 5 \\ 2 & 0 & 2 & 4 \\ 1 & 4 & 0 & 1 \\ 5 & 3 & 1 & 0 \end{matrix}$$



Ejemplo de funcionamiento:

Cada enlace tiene un peso asociado:

- **Nodos muy relacionados implican pesos** (valores de similitud) **altos y distancias** (valores de proximidad) **pequeñas**
- Si la matriz de adyacencia original contiene **grados de relación, se pueden invertir sus valores** para convertirlos en distancias



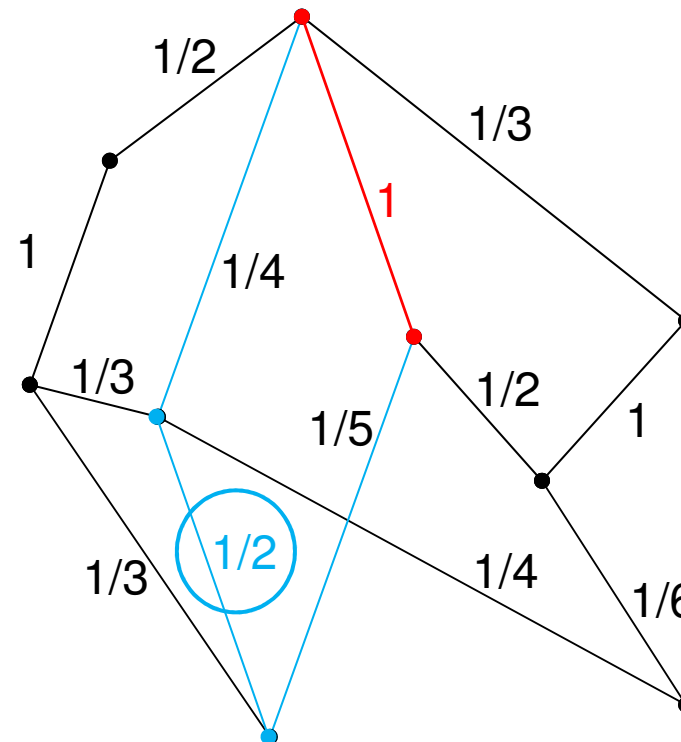
Ejemplo de funcionamiento:

Cada enlace tiene un peso asociado:

- **Nodos muy relacionados implican pesos** (valores de similitud) **altos y distancias** (valores de proximidad) **pequeñas**
- Si la matriz de adyacencia original contiene **grados de relación, se pueden invertir sus valores** para convertirlos en distancias

Tomamos $r=\infty$ y $q=n-1$ (8 en el ejemplo)

- La **distancia asociada a un camino** es la máxima distancia (el peso máximo) de los enlaces que lo componen
- Se considera cualquier camino sin ciclos



Ejemplo de funcionamiento:

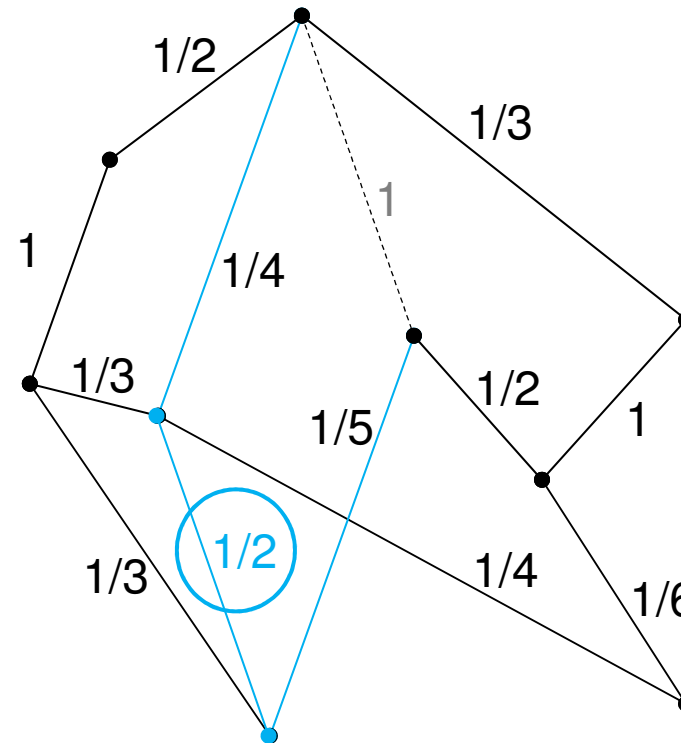
Cada enlace tiene un peso asociado:

- **Nodos muy relacionados implican pesos** (valores de similitud) **altos y distancias** (valores de proximidad) **pequeñas**
- Si la matriz de adyacencia original contiene **grados de relación, se pueden invertir sus valores** para convertirlos en distancias

Tomamos $r=\infty$ y $q=n-1$ (8 en el ejemplo)

- La **distancia asociada a un camino** es la máxima distancia (el peso máximo) de los enlaces que lo componen
- Se considera cualquier camino sin ciclos

Descartamos arcos con caminos alternativos de menor distancia



Ejemplo de funcionamiento:

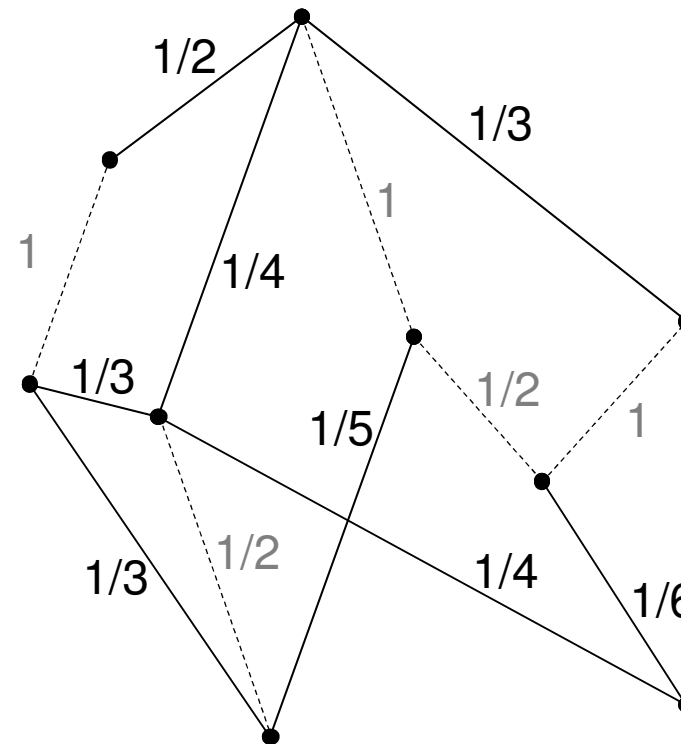
Cada enlace tiene un peso asociado:

- **Nodos muy relacionados implican pesos** (valores de similitud) **altos y distancias** (valores de proximidad) **pequeñas**
- Si la matriz de adyacencia original contiene **grados de relación, se pueden invertir sus valores** para convertirlos en distancias

Tomamos $r=\infty$ y $q=n-1$ (8 en el ejemplo)

- La **distancia asociada a un camino** es la máxima distancia (el peso máximo) de los enlaces que lo componen
- Se considera cualquier camino sin ciclos

Descartamos arcos con caminos alternativo de menor distancia



Ejemplo de funcionamiento:

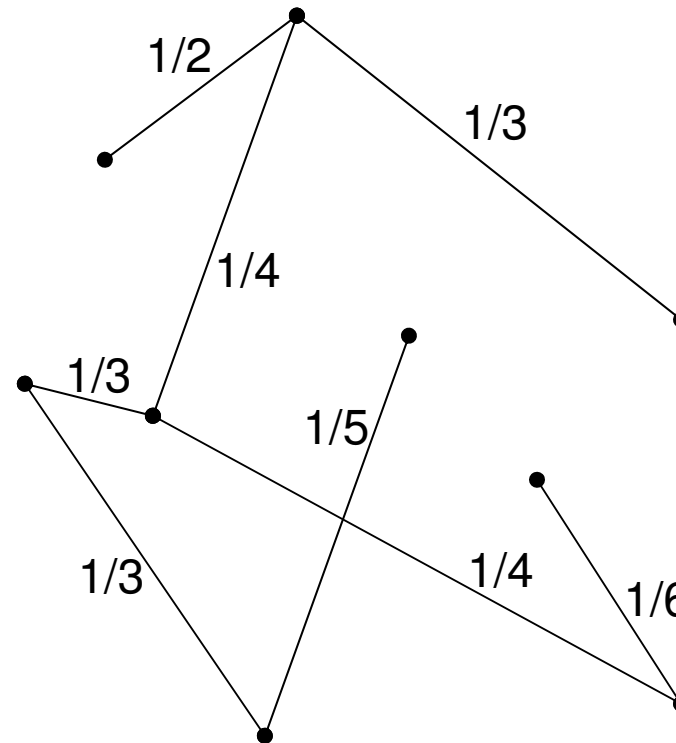
Cada enlace tiene un peso asociado:

- **Nodos muy relacionados implican pesos** (valores de similitud) **altos y distancias** (valores de proximidad) **pequeñas**
- Si la matriz de adyacencia original contiene **grados de relación, se pueden invertir sus valores** para convertirlos en distancias

Tomamos $r=\infty$ y $q=n-1$ (8 en el ejemplo)

- La **distancia asociada a un camino** es la máxima distancia (el peso máximo) de los enlaces que lo componen
- Se considera cualquier camino sin ciclos

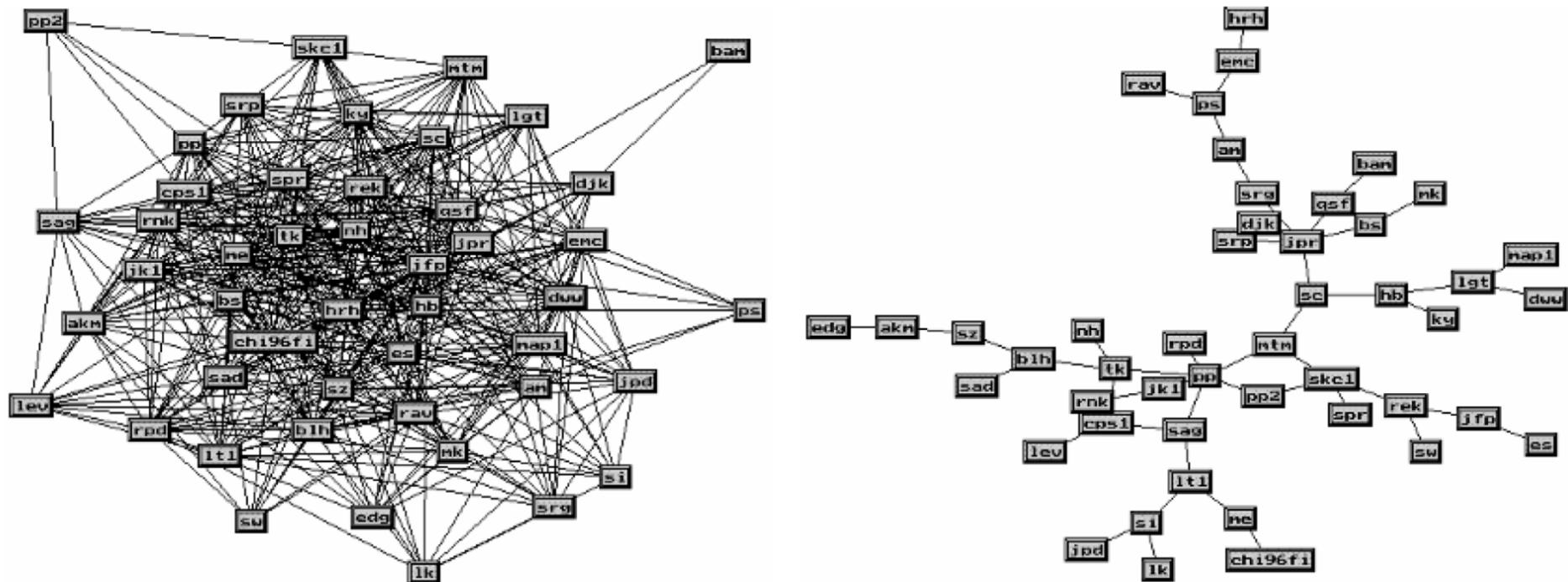
Descartamos arcos con caminos alternativo de menor distancia



EL ALGORITMO PATHFINDER DE PODA DE REDES

Ejemplos (7)

Ejemplo: PFNET($r=2, q=1$) (red original) vs. PFNET($r=2, q=n-1$)



EL ALGORITMO PATHFINDER DE PODA DE REDES

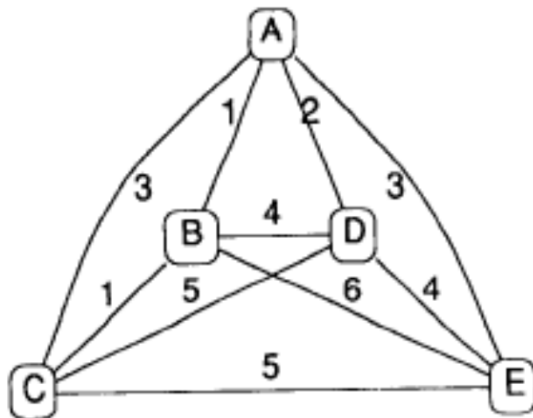
Ejemplos (8)

Ejemplo:

A Proximity Data (Adjacency Matrix)

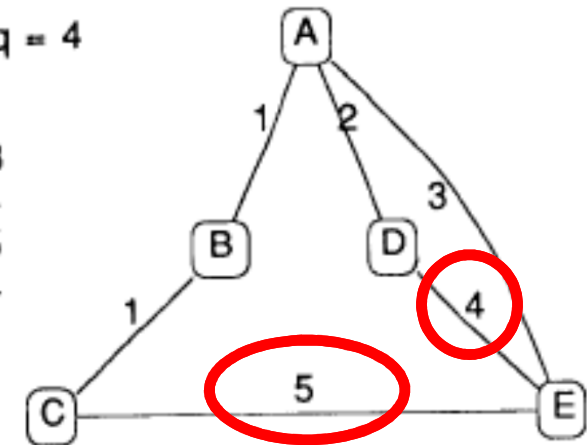
	A	B	C	D	E
A	0	1	3	2	3
B	1	0	1	4	6
C	3	1	0	5	5
D	2	4	5	0	4
E	3	6	5	4	0

DATANET



C Distance Matrix, $r = 1, q = 4$

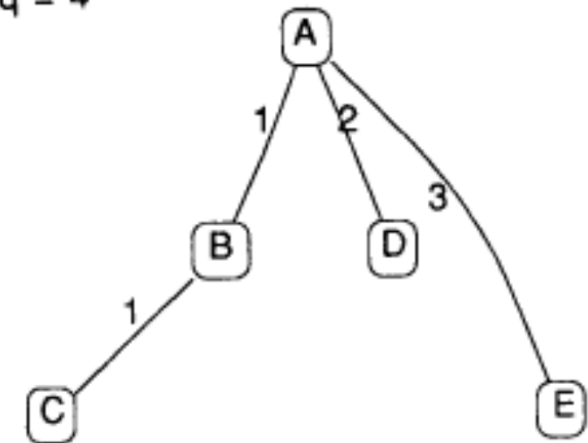
	A	B	C	D	E
A	0	1	2	2	3
B	1	0	1	3	4
C	2	1	0	4	5
D	2	3	4	0	4
E	3	4	5	4	0



PFNET($r = \infty, q = 4$)

B Distance Matrix, $r = \infty, q = 4$

	A	B	C	D	E
A	0	1	1	2	3
B	1	0	1	2	3
C	1	1	0	2	3
D	2	2	2	0	3
E	3	3	3	3	0



Ventajas de las PFNETs:

1. Constituyen un paradigma cuantitativo para el diseño de redes

En dominios en los que se dispone de una medida objetiva de similitud/distancia, proporcionan una representación única de la estructura subyacente que no es posible obtener con otros métodos de reducción de la dimensión:

- Modelan las relaciones asimétricas, lo que no es posible con MDS, que no permite mostrar ningún enlace (las relaciones entre objetos se representan por sus posiciones en la configuración espacial)
- Igualmente, representan de forma más precisa las relaciones locales que el MDS, el cual tiene que optimizar un criterio global

2. No sufren de las restricciones existentes en muchos algoritmos de clustering

3. Sólo muestran las relaciones más sobresalientes entre las componentes de la red

Dearholt, D., Schvaneveldt, R. 1990. Properties of pathfinder networks. En: R. Schvaneveldt (Ed.), Pathfinder associative networks: Studies in knowledge organization, pp. 1–30

Algoritmo básico para generar una red PFNET(r, q):

1. Define a network consisting of all nodes (concepts) N_i , but no links;
2. Order all elements e_{ij} of the E matrix in some nondecreasing order of their associated weights w_{ij} ;
3. Consider each e_{ij} , and include e_{ij} in the PFNET(r, q), if and only if e_{ij} provides a path from N_i to N_j which has a weight at least as small as the weight of any other path having no more than q links, using the r -metric to compute the weights of multiple-link paths.

Obviamente, su implementación directa daría lugar a un algoritmo muy ineficiente

Dearholt, D., Schvaneveldt, R. 1990. Properties of pathfinder networks. En: R. Schvaneveldt (Ed.), Pathfinder associative networks: Studies in knowledge organization, pp. 1–30

Los autores proponen el siguiente algoritmo como **método Pathfinder de poda de redes**:

1. Compute $W^{i+1} = W \odot W^i$, as follows: $w_{jk}^{i+1} = \text{MIN}((w_{jm})^r + (w_{mk}^i)^r)^{1/r}$, for $1 \leq m \leq n$.
2. Compute D^i , as follows: $d_{jk}^i = \text{MIN}(w_{jk}^1, \dots, w_{jk}^i)$, for $j \neq k$.
3. Iterate until W^q and D^q are computed.
4. Compare W^1 and D^q : all the links having the same values in these two matrices will belong to the final PFNET.

basado en el uso de dos tipos distintos de matrices de dimensión $n \times n$ auxiliares:

- W_{jk}^i , que almacenan los **costes mínimos** (los pesos de los caminos geodésicos) **para ir del nodo j al nodo k siguiendo exactamente i enlaces** (caminos de longitud i). W^1 es la matriz de pesos original de la red
- D_{jk}^i , que almacenan los **costes mínimos para ir del nodo j al nodo k siguiendo cualquier camino de la red compuesto por {1, ..., i} enlaces** (caminos de longitud menor o igual a i)

Funcionamiento:

Actually, W^1 is the weight matrix, and D^1 is identical to it, because it is the distance matrix for paths having one link. W^2 must be computed, however, and provides the minimum cost for paths between each pair of nodes having exactly two links. D^2 provides the lower cost of either one-link or two-link paths for each pair of nodes. Similarly, D^3 provides the lowest cost of paths having one, two, or three links, and finally, D^{n-1} (where n is the number of nodes) would provide the lowest cost of paths having any number of links (without cycles) between every pair of nodes. Thus the second step of the procedure assures that every link e_{ij} in $\text{PFNET}(r, q)$ provides a path between nodes N_i and N_j , which has a weight as small as any alternative path having from two to as many as q links. As each W^{i+1} and D^{i+1} are computed from W^i and D^i , links may be removed from the $\text{PFNET}(r, q)$, but they cannot be added. Those links which are removed are called *redundant* links, because they do not affect any of the distances between nodes (Schvaneveldt, Dearholt, & Durso, 1988).

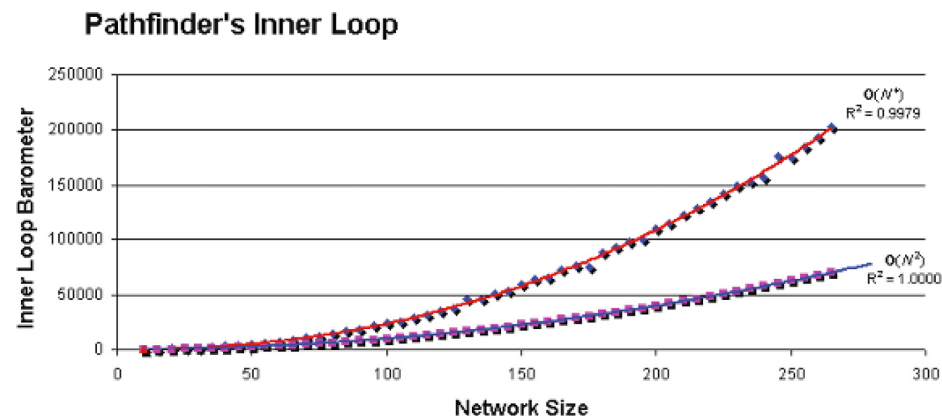
EL ALGORITMO PATHFINDER DE PODA DE REDES

Algoritmo Original (4)

Como puede observarse, el algoritmo está basado en **Programación Dinámica**. Para calcular su eficiencia, vemos que es necesario realizar q pasos para construir las q matrices W^i y D^i

Cada una de esas matrices almacena n^2 pesos por lo que es necesario un bucle de ese orden de complejidad para calcularlas en cada paso. Por último, es necesario un bucle adicional de n pasos para calcular cada componente de W^{i+1} (véase la línea 1 del algoritmo)

De este modo, el algoritmo tiene orden de complejidad $O(q \cdot n^3)$. Como el máximo valor posible para q es $n-1$, el algoritmo es de orden $O(n^4)$ en el peor caso (**¡costoso para redes grandes!**)



También es **costoso en términos de espacio de almacenamiento**. Dicho espacio es de orden $O(q \cdot n^2)$ ($O(n^3)$ en el peor caso) al tener que construir q matrices W^i y otras q matrices D^i

EL ALGORITMO PATHFINDER DE PODA DE REDES

Variantes

Se han propuesto **distintas variantes del algoritmo Pathfinder** para mejorar ambos aspectos, coste computacional y coste en espacio:

Name of the algorithm	Application domain	Time complexity (for $q = n - 1$)	Space complexity	Approach in algorithm theory
Original PF	Any valid values for q and r , (un-)directed graphs	$O(q \cdot n^3) = O(n^4)$	$2 \cdot n^2 + n$	Dynamic programming
Binary PF	Any valid values for q and r , (un-)directed graphs	$O(\log q \cdot n^3) = O(n^3 \cdot \log n)$	$2 \cdot n^2 + n$	Dynamic programming
Fast PF	Any valid values for $r, q = n - 1$, (un-)directed graphs	$O(n^3)$	$2 \cdot n^2 + n$	Dynamic programming
MST-PF (low-complexity)	$r = \infty, q = n - 1$, undirected graphs	$O(n^2 \cdot \log(n))$	$3 \cdot n^2 + n$	Greedy approach
MST-PF (practical)	$r = \infty, q = n - 1$, undirected graphs	$O(n^3)$	$3 \cdot n^2 + n$	Greedy approach

En realidad, puede considerar cualquier valor de q y el orden de complejidad es $O(q \cdot n^2)$

Las implementaciones están disponibles en:

<http://aquirin.ovh.org/research/mstpathfinder.html> ; <http://iv.slis.indiana.edu/sw/pfnet.html>

Guerrero-Bote, V., et al. 2006. Binary pathfinder: An improvement to the pathfinder algorithm. Information Processing and Management 42: 1484–1490

Binary Pathfinder es una variante del Pathfinder original que reduce el tiempo y el espacio requerido teniendo en cuenta los dos aspectos siguientes:

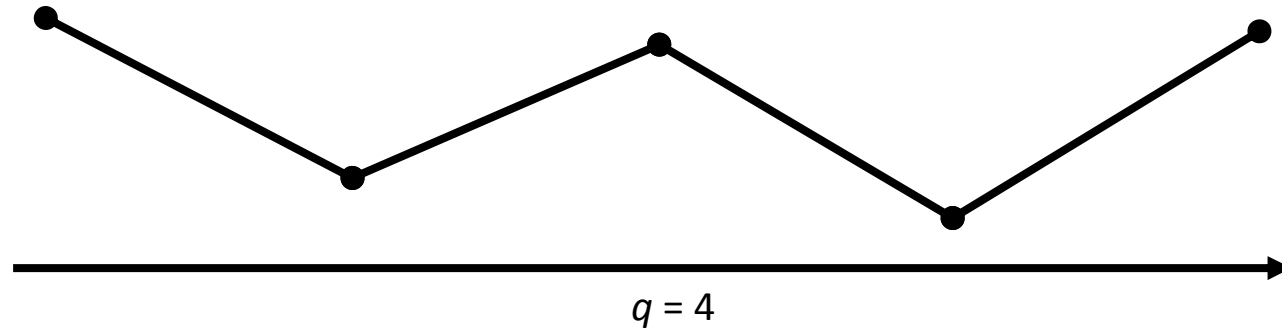
1. La única matriz de la serie D^i que es realmente necesaria es la última, D^q , que se compara con la matriz de pesos original W^1 para seleccionar los enlaces de la PFNET
2. Las matrices D^i pueden generarse a partir de las dos matrices anteriores en la serie al igual que se hace con las matrices W^i consecutivas: $D^{i+j} = D^i \otimes D^j$:

$$d_{kl}^{i+j} = \text{MIN}\{d_{kl}^i, d_{kl}^j, ((d_{km}^i)^r + (d_{ml}^j)^r)^{1/r}\} \quad ; \quad d_{kl}^1 = w_{kl}$$

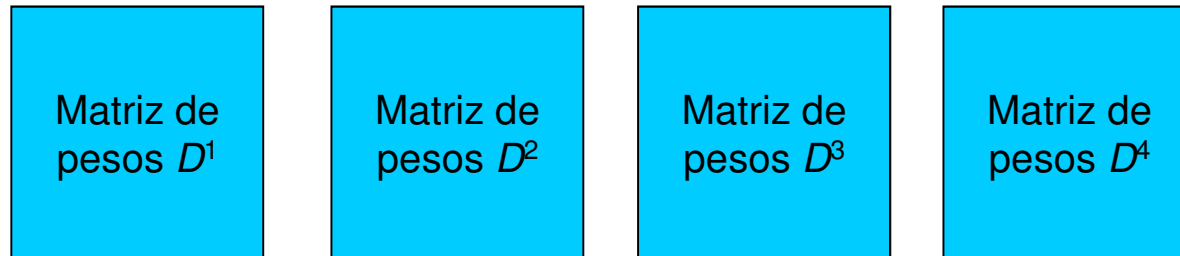
Basándose en esto, se propuso un nuevo algoritmo que **no necesita calcular cada matriz D^i** , $i=1, \dots, q$, sino que puede dar pasos más grandes. Inspirándose en el paso de números enteros a binario, el Binary Pathfinder calcula sólo $\log(q)$ matrices ($i=2^x$): $D^1, D^2, D^4, D^8, \dots$

EL ALGORITMO PATHFINDER DE PODA DE REDES

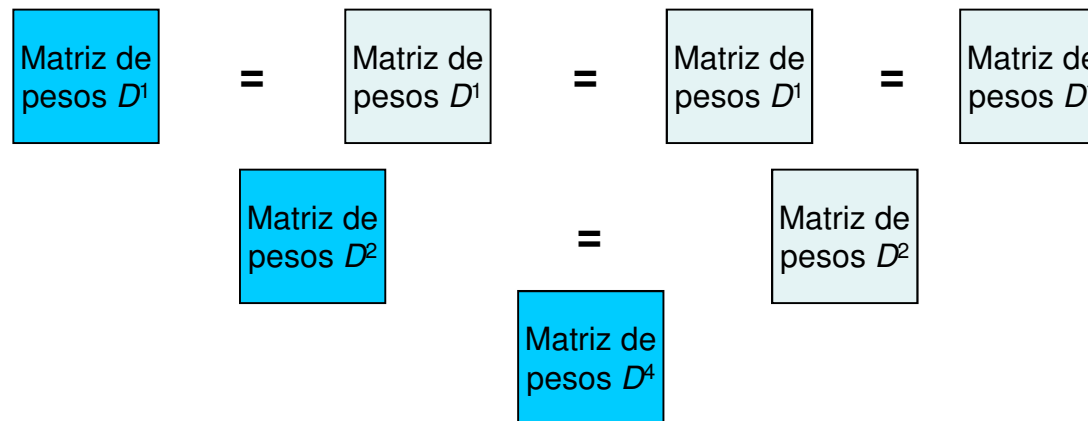
Binary Pathfinder (2)



Pathfinder original
 $O(n^4)$



Binary Pathfinder
 $O(n^3 \cdot \log n)$



1. $i = 1$; $nq = 0$; Generate $D^1 = W$; $D^q \leftarrow \infty$.
2. IF $(q \bmod 2 = 1)$ THEN Compute $D^q = D^q \odot D^1$.
3. $nq = 1$.
4. WHILE $(2 \cdot i \leq q)$
5. Compute $D^{2 \cdot i} = D^i \odot D^i$.
6. IF $((q - nq) \bmod (4 \cdot i) > 0)$ THEN
7. Compute $D^q = D^q \odot D^{2 \cdot i}$.
8. $nq = nq + 2 \cdot i$.
9. $i = 2 \cdot i$.
10. Compare W^1 and D^q : all the links having the same values in these two matrices will belong to the final PFNET.

El bucle principal reduce su número de pasos de q a $\log(q)$. Su orden de complejidad de tiempos pasa a ser $O(\log(q) \cdot n^3)$ ($O(\log(n) \cdot n^3)$ en el peor caso)

La reducción en espacio es aún mayor ya que sólo hacen falta **4 matrices**: 2 para calcular D^i en cualquier paso, otra para almacenar los valores finales de distancia, D^q , y una más W para guardar la matriz de pesos original; en lugar de las $2 \cdot q$ matrices W^i y D^i del Pathfinder original

Quirin, A., Cordon, O., et al. 2008. A new variant of the Pathfinder algorithm to generate large visual science maps in cubic time. *Information Processing and Management* 44: 1611–1623

La reducción de tiempos de Binary Pathfinder es importante pero no suficiente para ejecutar el algoritmo en tiempo real cuando se trabaja con redes grandes

Desde un punto de vista computacional, cuando $q=n-1$, Pathfinder sólo calcula una matriz D^{n-1} con los caminos mínimos entre cada par de nodos según la métrica r de Minkowski y compara sus valores con la matriz de pesos original para determinar qué enlaces pertenecen a la PFNET

En esa matriz, la desigualdad triangular se verifica siempre al trabajar con la matriz de caminos geodésicos de la red y **el problema se reduce al cálculo de dichos caminos mínimos**, un problema clásico y bien resuelto en Teoría de Grafos

Fast Pathfinder es una variante que implementa esta idea adaptando el algoritmo de caminos mínimos de Floyd-Warshall para que trabaje con la distancia de Minkowski

<http://www.cs.usfca.edu/~galles/visualization/Floyd.html>

1. $D \leftarrow W; PFNET \leftarrow \emptyset.$
2. FOR k from 1 to n DO
3. FOR i from 1 to n DO
4. FOR j from 1 to n DO
5. $d_{ij} = \text{MIN}\{d_{ij}, ((d_{ik})^r + (d_{kj})^r)^{1/r}\}.$
6. FOR i from 1 to n DO
7. FOR j from 1 to n DO
8. IF $(d_{ij} = w_{ij})$ THEN $PFNET \leftarrow PFNET \cup (i, j).$

Los pasos 1-5, del Floyd-Warshall, tienen orden $O(n^3)$. Los pasos 6-8 de selección de los enlaces de la PFNET, tienen orden $O(n^2)$. Por tanto, el orden final es $O(n^3)$

No se usan matrices temporales. El algoritmo sólo emplea 2 matrices: D^{n-1} y W

El algoritmo puede aplicarse para cualquier valor de q y r , basta con cambiar el primer bucle

Quirin, A., Cordón, O., et al. 2008. A Quick MST-Based Algorithm to Obtain Pathfinder Networks $(\infty, n-1)$. Journal of the American Society for Information Science and Technology 59: 1912–1924

La unión de todos los MSTs extraídos de una red equivale a su PFNET($r=\infty, q=n-1$)

MST-Pathfinder es una última variante de Pathfinder que genera la PFNET($r=\infty, q=n-1$) de una red en el mismo orden de eficiencia que los algoritmos de MST ($O(n \cdot \log n)$) **generando todos los MSTs de la red original y uniéndolos**

Se basa en un enfoque *greedy* (más rápido) a costa de restringir los posibles valores de r y q a sólo ∞ y $n-1$. Aún así, esa es la configuración más empleada, al ser la de mayor poder de poda

MST-Pathfinder puede basarse en cualquiera de los dos algoritmos clásicos de generación de MSTs, Kruskal y Prim

La adaptación de Kruskal es más sencilla y más eficiente en redes dispersas

1. Define a tree $T = \emptyset$.
2. Define $V[G]$, the set of the nodes of the network G .
3. For each node $v \in V[G]$
4. CREATE-CLUSTER(v).
5. Create F , a set of all the links of G sorted by their weights.
6. FOR each link $e(u, v) \in F$
7. IF CLUSTER(u) \neq CLUSTER(v), THEN
8. $T = T \cup \{e(u, v)\}$.
9. MERGE-CLUSTER(u, v).
10. Return T .

1. Define a tree $T = \emptyset$.
2. Define $V[G]$, the set of the nodes of the network G .
3. Define W , the matrix of the costs for each link of G .
4. For each node $v \in V[G]$
5. CREATE-CLUSTER(v).
6. Create F , a set of all the links of G sorted by their weights.
7. FOR each link $e(u, v)$ remaining in F
8. $H = \emptyset$.
9. FOR each link $e(u', v')$ remaining in F where $w(u, v) = w(u', v')$
10. $F = F - \{e(u', v')\}$.
11. IF CLUSTER(u') \neq CLUSTER(v'), THEN
12. $T = T \cup \{e(u', v')\}$.
13. $H = H \cup \{e(u', v')\}$.
14. FOR each link $e(u', v') \in H$
15. MERGE-CLUSTER(u', v').
16. Return T .

Si todos los pesos de la matriz inicial son distintos, sólo hay un MST. En otro caso, existen varios asociados a los **enlaces especiales** (con los mismos pesos)

MST-Pathfinder detecta esos enlaces, los almacena en un conjunto temporal H y los incluye en el MST sólo cuando no afectan a la detección de otros enlaces especiales

Al contrario que en los algoritmos anteriores, **no es necesario comparar pesos entre distintas matrices para construir la PFNET** ya que los arcos que la componen ya están en los MSTs

El algoritmo necesita $O(|E| \cdot \log|E|)$ operaciones para ordenar los enlaces por peso, donde $|E|$ es el número de enlaces de la red

Para saber a qué cluster pertenece cada nodo, se usa una estructura de datos específica, un *conjunto disjunto*, cuyas operaciones también están acotadas superiormente por ese orden

En redes densas, $|E|=n^2$, y el algoritmo tiene orden $O(n^2 \cdot \log(n))$ en el peor caso

Este tiempo es teórico. En la práctica se usa una estructura de datos más simple que hace que el algoritmo sea de orden $O(n^3)$ pero que funcione más rápido

Es almacenar 3 listas de enlaces con sus pesos (F , T y H). El índice del cluster se almacena con un único atributo adicional. Por tanto, el espacio requerido es $3 \cdot n^2 + n$

Eso sí, **el algoritmo no funciona en redes dirigidas** por el proceso de ordenación

TABLE 1. Comparison of the run time (expressed in seconds) of all the algorithms for the random matrices case studies.

#	#Nodes	#Links	Original PF	Binary PF	Fast-PF	MST-PF (low-complexity)	MST-PF (practical)
1	100	9.90E+03	1.55	0.183	0.00925	0.0021	0.00208
2	200	3.98E+04	23.59	1.76	0.0702	0.0101	0.0101
3	300	8.97E+04	181	8.98	0.238	0.0266	0.0264
4	400	1.60E+05	604	24.56	0.585	0.0537	0.0533
5	1000	9.99E+05	>3600	>3600	10.01	0.629	0.629
6	10000	1.00E+08	>3600	>3600	>3600	128.31	127.62