Análisis de Redes Sociales

Guillermo Jiménez Díaz (gjimenez@ucm.es)
Alberto Díaz (albertodiaz@fdi.ucm.es)

13 de noviembre de 2014

Prefacio

Estos son los apuntes de la asignatura Análisis de Redes Sociales, impartida en la Facultad de Informática de la Universidad Complutense de Madrid por los profesores Guillermo Jiménez Díaz y Alberto Díaz, del Departamento de Ingeniería del Software e Inteligencia Artificial.

Este material ha sido desarrollado a partir de distintas fuertes, destacando como referencia principal el libro *Network Science* de Laszlo Barabasi, el material de la asignatura *Social Network Analysis*, impartido por Lada Adamic a través de Coursera, y las transparencias de la asignatura Redes y Sistemas Complejos, creadas por Óscar Cordón García de la Universidad de Granada.

Tema 6: Estructura de comunidades

5.1 Introducción

5.1.1 Concepto de comunidad

Las redes complejas tienden a mostrar una estructura de comunidades. Esta propiedad suele darse como consecuencia de la heterogeneidad global y local de la distribución de los enlaces en un grafo. A menudo encontramos una alta concentración de enlaces en ciertas regiones del grafo, denominadas comunidades, y una baja concentración de enlaces entre esas regiones. Las comunidades, también conocidas como módulos o clusters, se definen de forma sencilla como grupos de nodos similares. A partir del concepto de densidad de la red, las comunidades pueden definirse como grupos de nodos densamente conectados que presentan conexiones dispersas entre sí.

5.1.2 Caso ideal

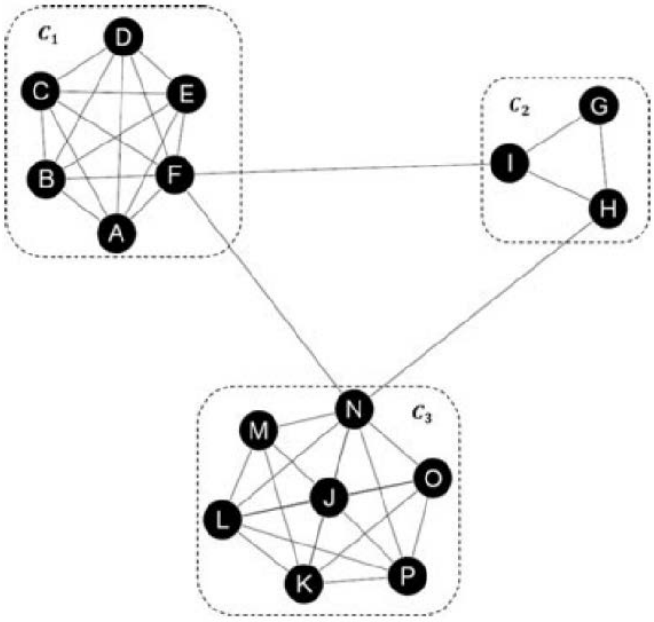


Figure 1: Caso ideal de estructura de comunidades

En esta red hay 3 comunidades: $C_1, C_2 y C_3$. Cada comunidad está formada por un grafo completo (un clique) de tamaño variable ($C_1 = K_6, C_2 = K_3, C_3 = K_7$). La densidad de enlaces entre comunidades es muy baja. Los pocos enlaces que existen son puentes.

5.2 ¿Por qué detectar comunidades?

En la vida real existen muchos ejemplos de grupos que forman comunidades en redes complejas:

- Sociedades: las personas tienen una tendencia natural a formar grupos (familias, círculos de amigos, grupos profesionales o religiosos, ciudades, naciones, etc.)
- Empresariales/Económicas: compañías, clientes, etc.
- Biología: p.ej. redes metabólicas. En redes de interacción de proteínas encontramos grupos de proteínas con funciones similares dentro de la célula.
- Internet: comunidades virtuales (Facebook, Twitter, etc.), grupos de páginas web relacionadas, etc. (útiles para sistemas de recomendaciones)
- y muchos ámbitos más...

5.2.1 Ejemplo email spectroscopy

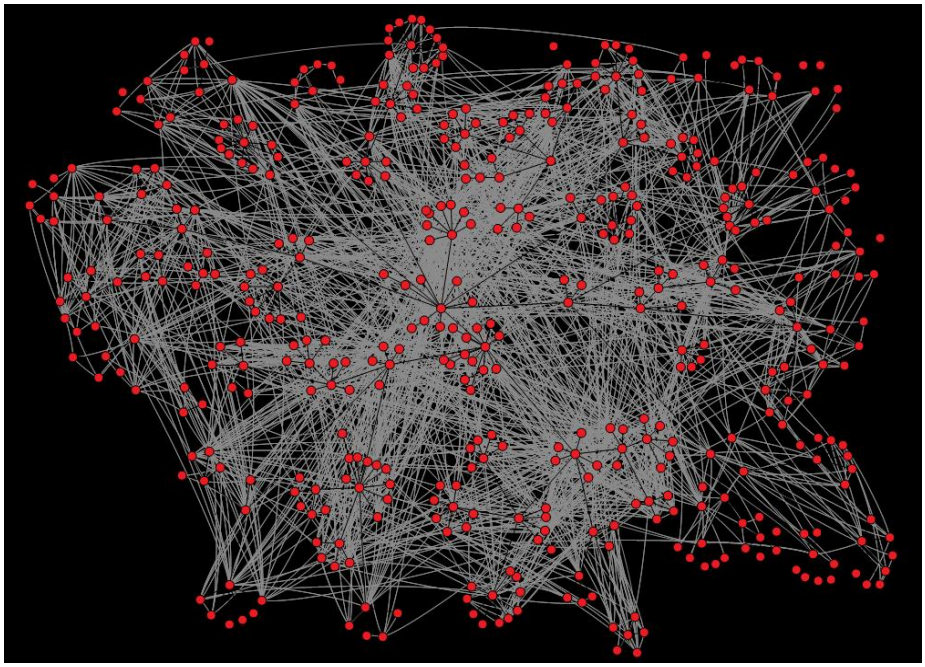


Figure 2: Grafo con conexiones por correos electrónicos (gris) y estructura de comunicación entre unidades organizativas en una empresa (HP)

En la Figura aparece un grafo con las conexiones entre las personas de una empresa (HP) según los correos electrónicos que intercambian. Detectar comunidades en este grafo nos da la estructura de colaboración entre las personas que trabajan en la empresa.

En gris aparecen los enlaces debidos a los correos y en negro aparecen las conexiones debidas a la estructuración de la empresa desde el punto de vista de la comunicación entre personas (reporting structure).

En este estudio, se obtuvieron las comunidades presentes en el grafo de correos y se compararon con la estructura de comunicación de la empresa. Se encontró que en la mayoría de los casos los grupos de personas coincidían. Esto es, las personas que trabajan en el mismo departamento, con los mismos objetivos y tareas, tienden a comunicarse más entre ellos que con el resto.

Esto supone una validación de que dicha estructura es útil, de tal forma que si no hubiese salido de forma similar la empresa podría replantearse algún tipo de re-organización de sus departamentos y personas.

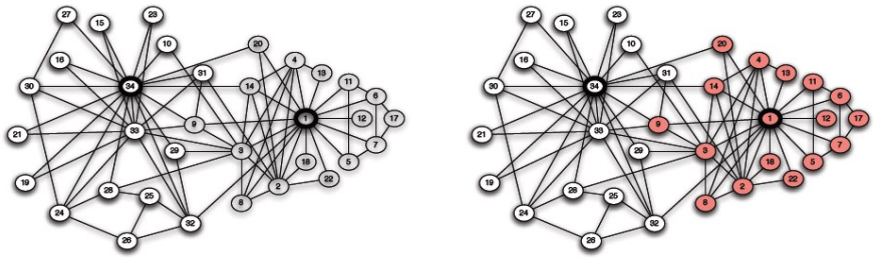
De hecho, si la empresa no tuviera una estructura organizativa, las comunidades obtenidas en el análisis de la red de correos podrían ser una sugerencia sobre

una posible estructura organizativa de la que partir.

Otra cosa que fue detectada en este estudio es que había unas pocas comunidades que incluían personas de varias unidades organizativas de la empresa. Cuando los autores del estudio entrevistaron a las personas incluidas en estas comunidades resultó que dichas personas realizaban tareas que estaban relacionadas con varias partes de la organización. Por ejemplo, en el trabajo asociado con impresoras con funcionalidad de impresión en red es necesario la gente que trabaja en hardware de impresoras y gente que trabaja en manejo de la red, que pertenecen a 2 unidades organizativas distintas.

De esta forma, las comunidades detectadas dan una visión más real de la estructura de colaboración dentro de la gente que trabaja en la empresa, más allá del simple hecho de trabajar en diferentes unidades organizativas.

5.2.2 Ejemplo: El club de karate de Zachary



(a) *Karate club network* (b) *After a split into two clubs*

Figure 3: Red del club de karate de Zachary.

Fuente: Zachary. An information flow model for conflict and fission in small groups. J. Anthropol. Res. 33: 452-473 (1977)

Zachary, un sociólogo, estudió las interacciones entre las personas de un club de karate universitario. Creó una red social donde enlazaba a las personas que hablaban regularmente y a quien le gustaba cada persona.

Mientras estaba estudiando las interacciones en el club, hubo un problema interno y el club se dividió en 2 clubs.

La pregunta es, dada la información de la red (quien habla con quien y a quien le gusta cada persona), se puede predecir como se va a dividir el club?

Esto es un tipo de test de prueba de si el algoritmo de detección de comunidades es bueno o no.

Este tipo de información podría ser útil para el análisis de este tipo de casos en cualquier red social.

5.2.3 Ejemplo: Red del aserradero

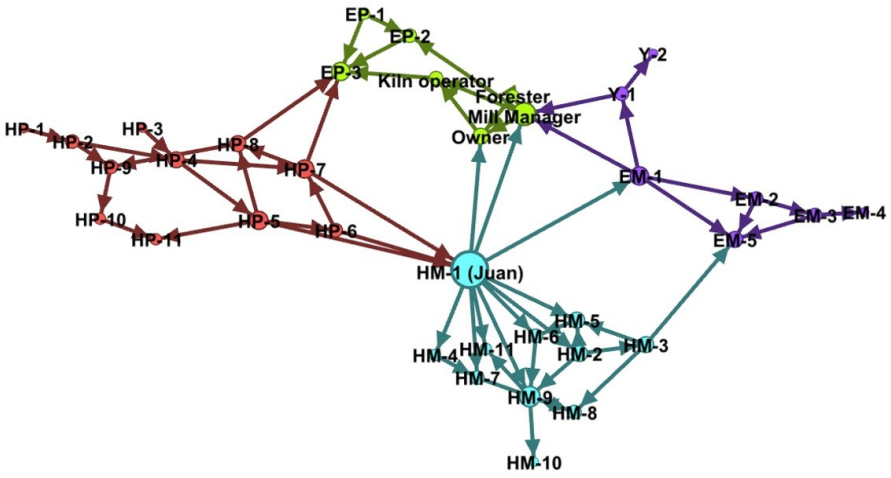


Figure 4: Red del aserradero.

Fuente: Exploratory Social Network Analysis with Pajek

En esta red aparecen los trabajadores de un aserradero. Los trabajadores se pueden clasificar en varias categorías:

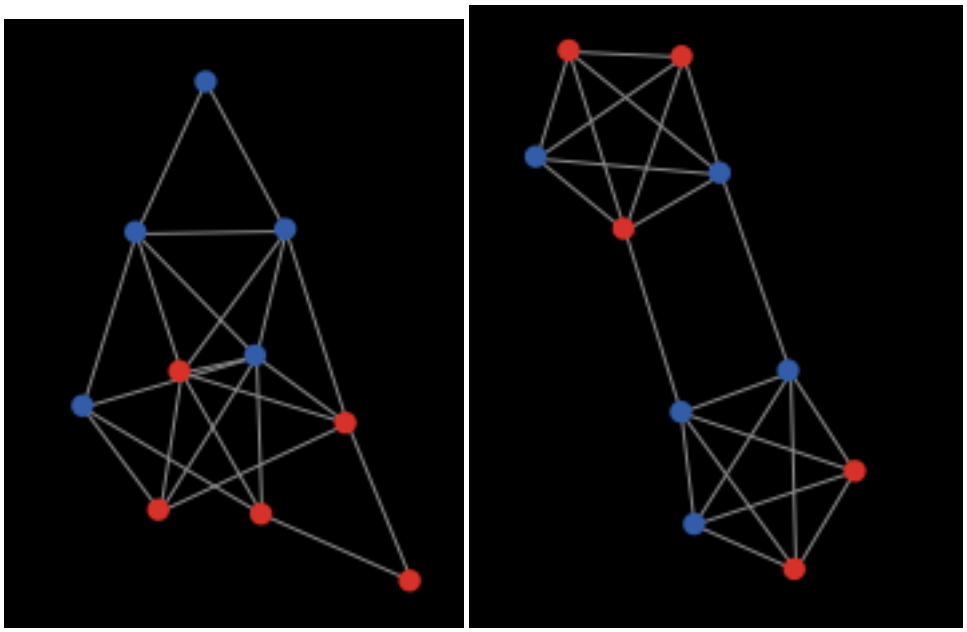
- E = ingleses, H = hispanos
- P = planificación, M = corte, Y = patio

La dirección del aserradero estaba teniendo dificultades para persuadir a los trabajadores para que adoptaran un nuevo plan que redundaría en un beneficio para todos los empleados. En concreto, los trabajadores hispanos eran los más reticentes a aceptar.

La dirección contrató a un sociólogo que diseñó una red que reflejaba con qué compañeros hablaba cada persona habitualmente. El sociólogo recomendó a la dirección que hablaran con Juan y que le pidieran que hablara con el resto de trabajadores hispanos. Fue un éxito, rápidamente todos estaban de acuerdo con el nuevo plan.

5.2.4 Formación de opiniones y estructura de comunidades

http://www.ladamic.com/netlearn/NetLogo502/OpinionFormationModelToy.html



Ejecución en NetLogo

- Se pueden crear 2 tipos de redes: una aleatoria y otra con estructura de 2 comunidades (comunnity? on/off).
- Con Update Opinion cada nodo adopta la opinión mayoritaria de sus vecinos (una opinión aleatoria, en caso de empates).
- En el modelo netlogo, si la mayoría de los vecinos de un nodo tienen opinión roja, entonces el nodo cambia a opinión roja, si no la tenía antes. Si hay igual número de rojos que de azules entonces elige aleatoriamente si cambia o no.
- Si cambia de opinión cambia de circulo a cuadrado.

PREGUNTA: Ejecutar alternativamente la configuración de dos comunidades y la de Erdos-Renyi. ¿Cuál puede mantener opiniones divergentes cuando se itera la actualización de opiniones?

a) Sólo la de Erdos-Renyi

b) Sólo la de dos comunidades

c) Ambas

Podemos repetir la actualización de opiniones muchas veces y entonces, algunas veces, aunque no siempre, diferentes partes de la red mantendrán una opinión diferente.

Si tenemos dos comunidades separadas, una de las comunidades puede mantener su opinión aunque el resto de la red piense otra cosa. Esto es porque los individuos de la comunidad tienen más enlaces entre ellos que con el resto de individuos de fuera de la comunidad.

Si cada nodo adopta la opinión de la mayoría de sus vecinos, es posible formar opiniones distintas en subgrupos cohesivos distintos. Hay más uniformidad dentro de un grupo cohesivo.

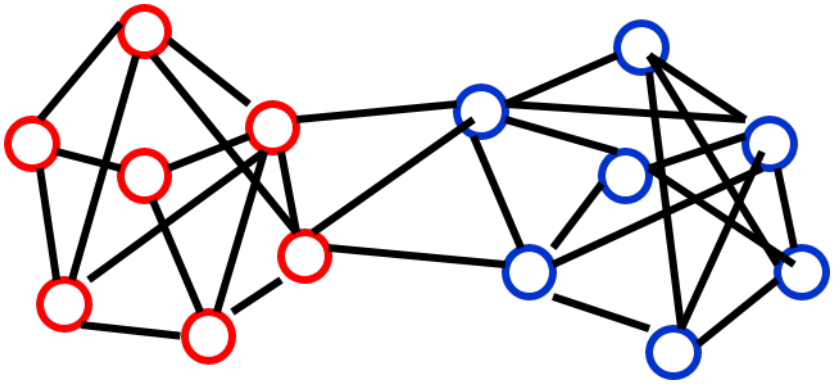


Figure 5: Red con 2 comunidades.

5.2.5 Mapa de la ciencia de citaciones de 6000 revistas

Fuente: Rosvall y Bergstrom. Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. USA 105: 1118-1123 (2008)

En esta investigación se buscaban patrones de acceso a repositorios online de artículos científicos, por ejemplo, JSTOR.

En este primer mapa se muestra los enlaces entre distintas áreas de investigación. De tal forma que si uno accede a un artículo en un área, los enlaces muestran qué otras áreas de investigación es probable que uno explore en la misma sesión. Posteriormente se han agrupado en comunidades.

En el segundo mapa se puede ver con más detalle, por ejemplo, plant biology está relacionado con ecology, biodiversity y environmental science.

Además si se colapsan las diferentes comunidades en metanodos, se puede ver que metanodos están relacionados con otros.

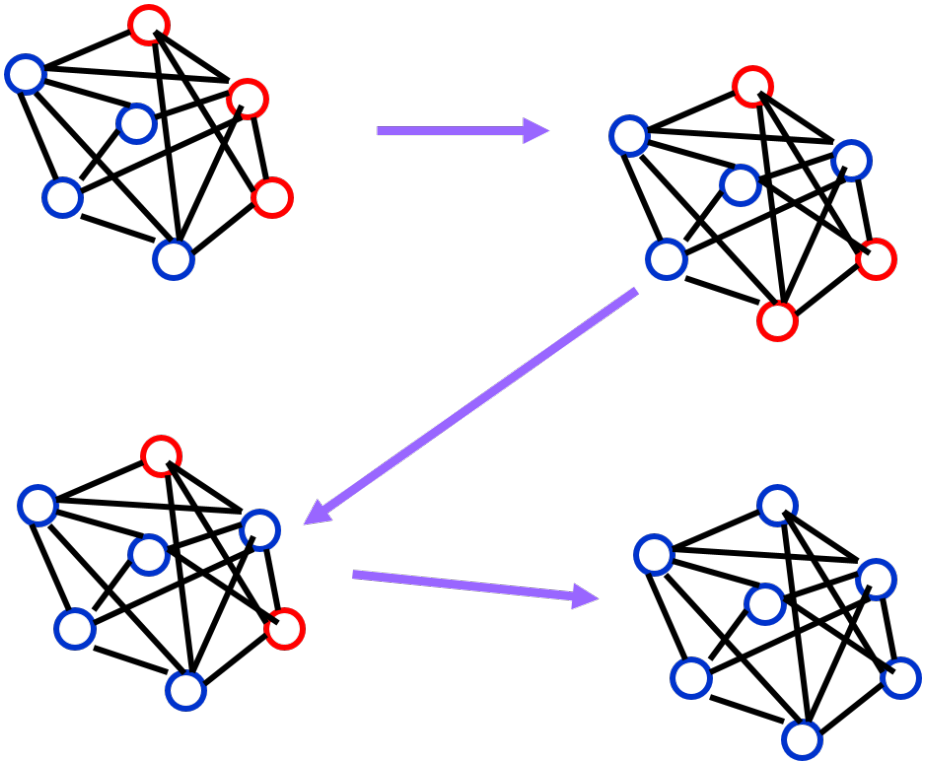


Figure 6: Red aleatoria.

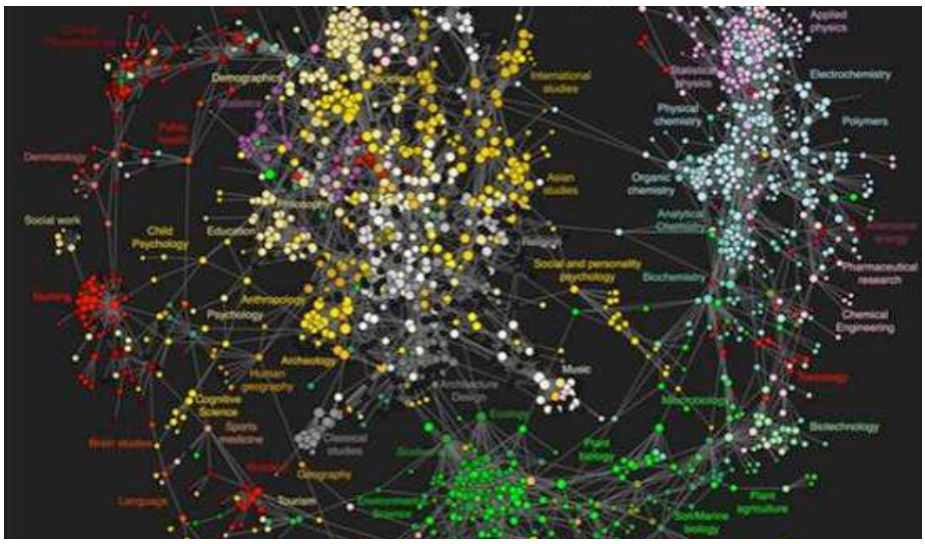


Figure 7: Mapa de la ciencia de citaciones de 6000 revistas.

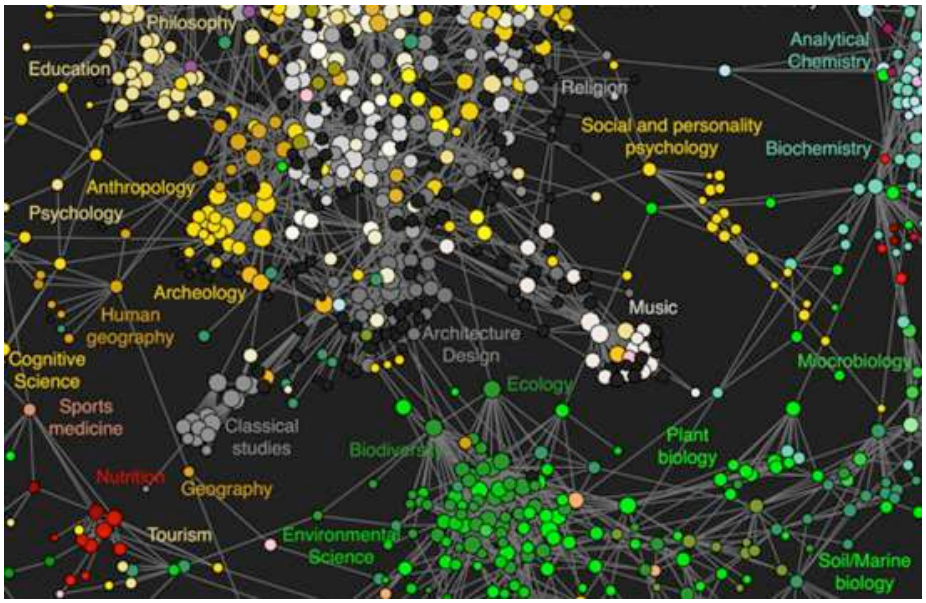


Figure 8: Mapa de la ciencia de citaciones de 6000 revistas.

5.3 Criterios estructurales que identifican una comunidad en una red

En este apartado vamos a ver distintas formas de considerar un conjunto de nodos como una comunidad. Se pueden utilizar distintos criterios:

- Mutualidad completa (cliques): El grupo es un subgrafo completo (todo el mundo conoce a todo el mundo en el grupo).
- Frecuencia de enlaces entre los miembros (k-cores): Todos los miembros del grupo tiene enlaces al menos a otros k miembros. Cada miembro conoce al menos k miembros dentro de su comunidad.
- Alcanzabilidad/cercanía entre los miembros (n-cliques): Los individuos del grupo están separados por un máximo de n saltos. Se puede llegar de un miembro a otro del grupo en un número pequeño de saltos.
- Comparación de la cohesión interna y externa del grupo (p-cliques): frecuencia relativa de enlaces entre los miembros del grupo en comparación con la de los no miembros. Es suficiente con conocer una cierta proporción de individuos dentro de la comunidad.

Fuente: Wasserman y Faust. Social Network Analysis. Cambridge University Press; 1994

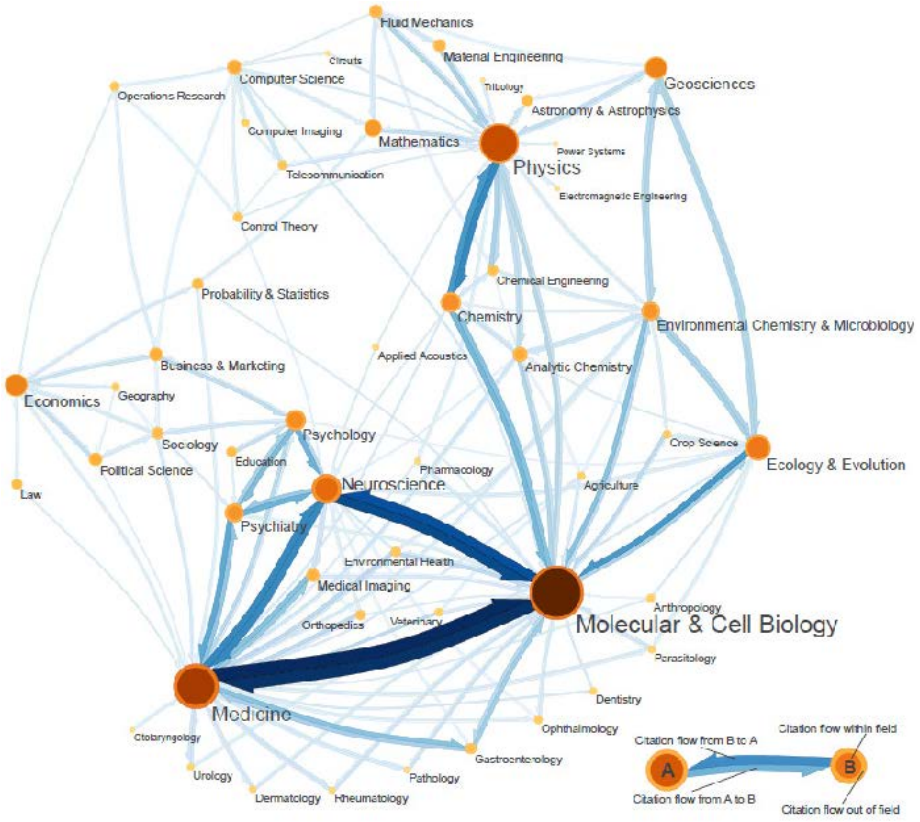


Figure 9: Mapa de la ciencia de citaciones de 6000 revistas.

5.3.1 Identificación de cliques

Todos los miembros del grupo tienen enlaces al resto dentro de la misma comunidad. Todo el mundo conoce a todo el mundo en la misma comunidad. Un clique es un subgrafo completo

Los triángulos son los cliques más básicos, los de mayor dimensión son menos frecuentes.

Los cliques pueden estar solapados

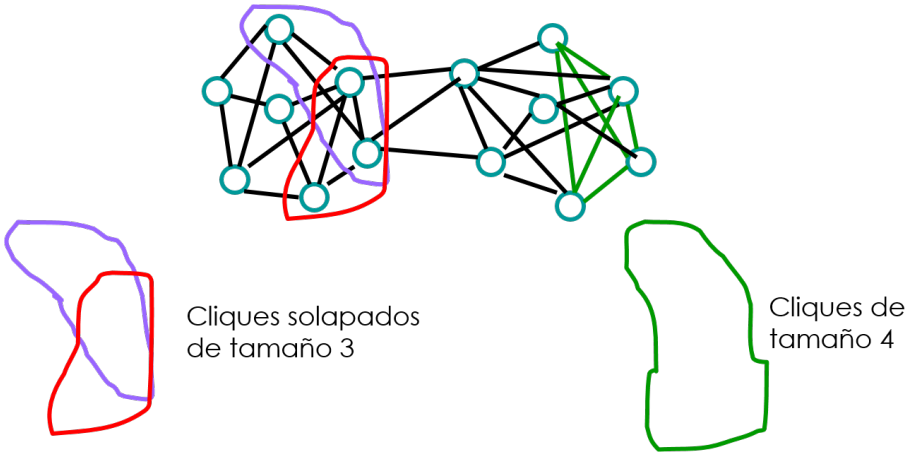


Figure 10: Ejemplos de cliques.

En la Figura hay 2 cliques, cada uno con 3 miembros cada uno. Estos 2 cliques están solapados. Por otro lado, hay un clique con 4 miembros.

Modelo Netlogo: http://www.ladamic.com/netlearn/nw/Cliques.html.

- setup con community on/off, layout para visualizar mejor la red.
- al pulsar en biggest maximal clique se obtiene el clique de mayor tamaño.

Comparar la configuración de red aleatoria ER con la de estructura de comunidades (son las mismas que en el modelo de formación de opiniones).

PREGUNTA: ¿Cuál de los dos redes tiene un clique máximo de mayor tamaño?

La red con 2 comunidades tiene un clique de tamaño 5. En la red aleatoria depende de la ejecución, el clique mayor suele tener tamaño 4, pero puede tener tamaño 3 o 5 en algunos casos.

En general, las redes con estructura de comunidades tienden a tener cliques de mayor tamaño que las redes aleatorias equivalentes.

5.3.2 Problemas en la identificación de cliques

Es un problema NP-completo.

No son robustos. * Un solo enlace faltante puede descalificar un clique, haciendo que el grupo no sea considerado una comunidad. Demasiado restrictivo.

No son interesantes * Todo el mundo está conectado entre sí. * No hay estructura central densamente conectada y un conjunto de individuos más periféricos. Sólo hay una estructura central totalmente conectada. * Las medidas de centralidad no dan información. Todos los individuos tienen el mismo valor en dichas medidas.

Por otro lado, el solapamiento de cliques puede ser más relevante que su propia existencia. Volveremos sobre esto al final del tema.

5.3.3 Identificación de k-cores

Cada nodo de un grupo está conectado con otros k nodos de dicho grupo. No hacen falta conexiones con todos.

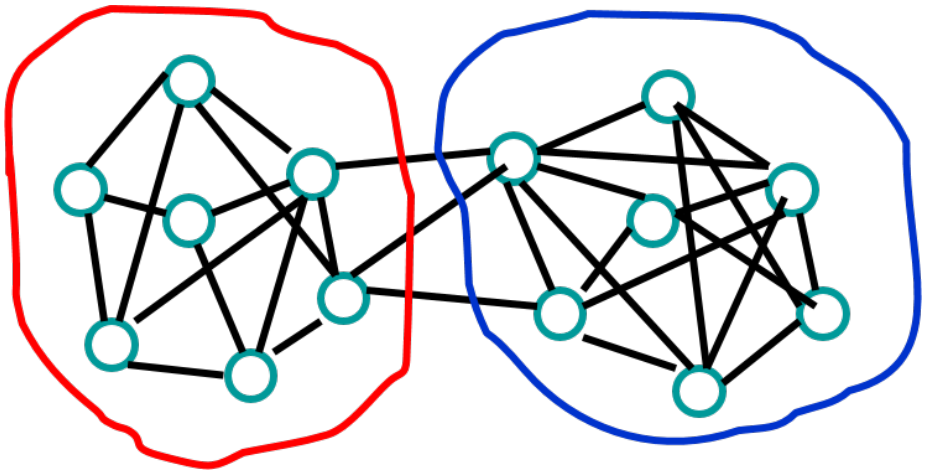


Figure 11: Ejemplos de k-cores.

Pregunta:

- ¿Cuál es el valor de k para el k-core marcado en rojo?
- ¿y para el azul?

Cada nodo de un grupo está conectado con otros k nodos de dicho grupo:

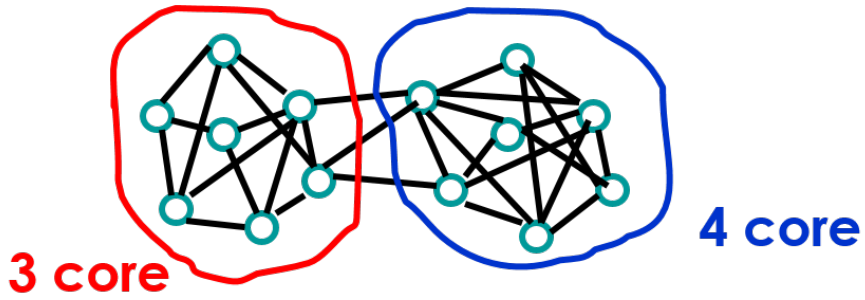


Figure 12: Ejemplos de k-cores.

- rojo: 3-core. Cada nodo está conectado con al menos otros 3 nodos distintos dentro del mismo grupo.
- azul: 4-core. Cada nodo conectado con al menos otros 4 nodos del mismo grupo.

5.3.4 Problemas

Aun así, es una estructura demasiado restrictiva como requisito para identificar comunidades naturales.

En la siguiente Figura se observa un una red con un nodo que debería pertenecer al grupo del 4 core, pero como sólo tiene 2 aristas no puede pertenecer al grupo, necestiaría tener al menos 4 aristas. En realidad habría un 2-core que englobaría a todos los nodos y por tanto no detectaría 2 comunidades.

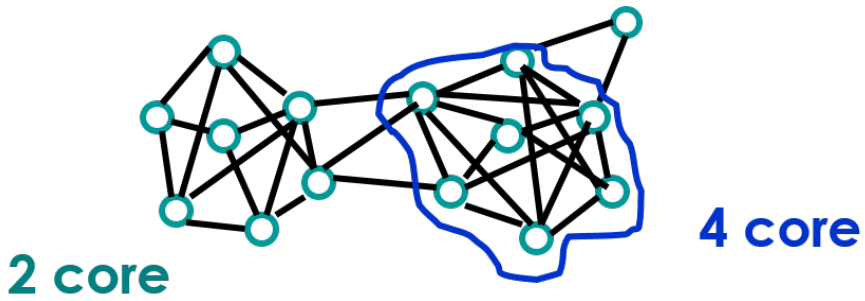


Figure 13: Ejemplos de k-cores.

5.3.5 Identificación de n-cliques basada en alcanzabilidad y diámetro

En este caso lo que se tiene es el flujo de información en la red, es decir, lo importante es que se puede llegar de un nodo a otro con un número pequeño de saltos, por ejemplo, si se puede llegar con 2 saltos de un miembro a todos los demás miembros de un grupo entonces ese grupo es un 2-clique.

Dicho de otra forma, la máxima distancia entre cualesquiera dos nodos del grupo es n.

En la Figura aparecen dos 2-cliques. En cada 2-clique la máxima distancia entre cualesquiera dos nodos del grupo es 2.

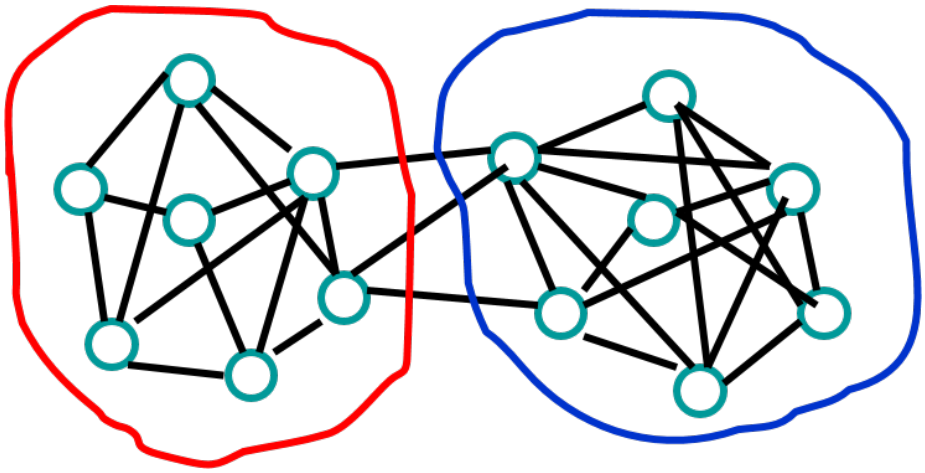


Figure 14: Dos 2-cliques.

5.3.6 Problemas

El n-clique puede estar desconectado (los caminos pueden pasar por nodos que no estén en el grupo). En la Figura se puede observar que el nodo rojo estaría fuera del 2-clique, pero sin embargo el camino rojo uniría 2 nodos del 2-clique. Esto está asociado a que el diámetro (camino más largo entre cualesquiera 2 nodos) es igual a 3, que es mayor que n, igual a 2.

5.3.7 Identificación de p-cliques

Particionamiento de la red en clusters en los que los nodos tienen como mínimo una proporción pÎ[0,1] de vecinos dentro del grupo. Tiene en cuenta la cohesión de enlaces en el grupo.

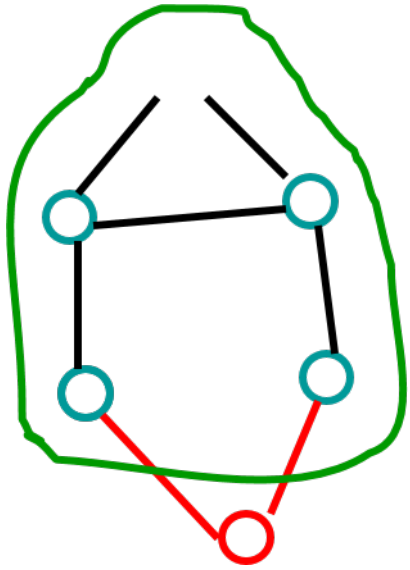


Figure 15: Ejemplo.

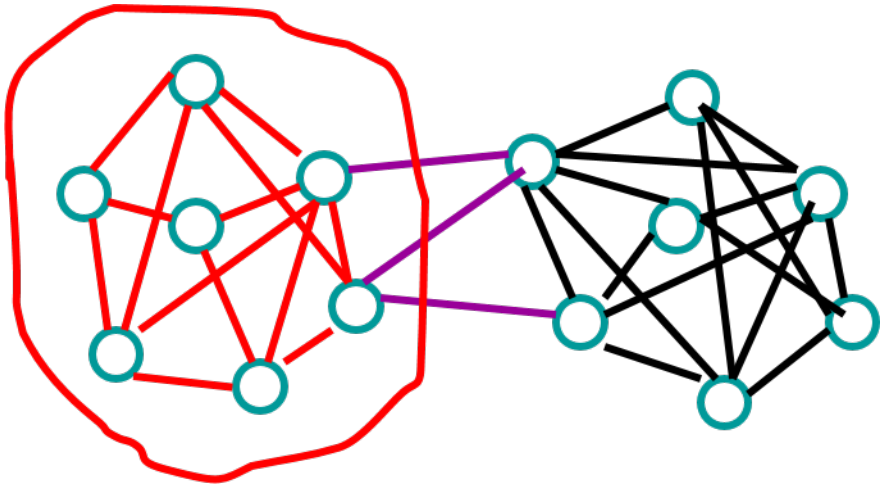


Figure 16: Ejemplo.

5.3.8 Cohesión en redes dirigidas y ponderadas

Se aplica obteniendo componentes fuertemente conexas. Es un proceso costoso. Es conveniente podar enlaces antes de aplicarlo.

Se podrían podar enlaces de 2 formas:

- Teniendo en cuenta reciprocidades. En el ejemplo de la cena de las 26 chicas se podrían eliminar los enlaces que no correspondieran a elecciones recíprocas. Esto reduciría la red y permitiría ver comunidades más coherentes.
- Podar enlaces basándose en un umbral sobre los pesos asignados a las aristas. Si la red representara frecuencia de cuanto habla una persona con otra, podría ser más significativo eliminar enlaces a partir de un cierto umbral de tiempo a la hora de detectar comunidades.

5.3.9 Ejemplo

Ejemplo: Blogs políticos (29 Ago-15 Nov 2004)

A. Todas las citas entre blogs en los dos meses anteriores a la elección de 2004 B. Citas entre blogs con al menos 5 citas en ambas direcciones C. Poda más restrictiva incluyendo enlaces con al menos 25 citas combinadas

Sólo un 15% de las citas construyen comunidades.

Fuente: Adamic y Glance. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. Proc. LinkKDD2005

5.4 Enfoques para detección de comunidades

Según Newman y Girvan (2004), existen dos líneas de investigación principales para el descubrimiento de comunidades en redes complejas:

1. Particionamiento de grafos: tiene su origen en Informática, en el campo de la computación distribuida. Busca la mejor forma de asignar tareas a procesadores para minimizar las comunicaciones entre ellos.
2. Modelado de bloques (también llamado clustering jerárquico o detección de la estructura de comunidades): se origina en Sociología. Está motivado por el descubrimiento de grupos en una sociedad para facilitar el análisis de fenómenos sociales.

En cualquier caso, el procedimiento implica dividir el grafo original en un conjunto de subgrafos disjuntos mediante la optimización de una función objetivo (p.ej. la modularidad)

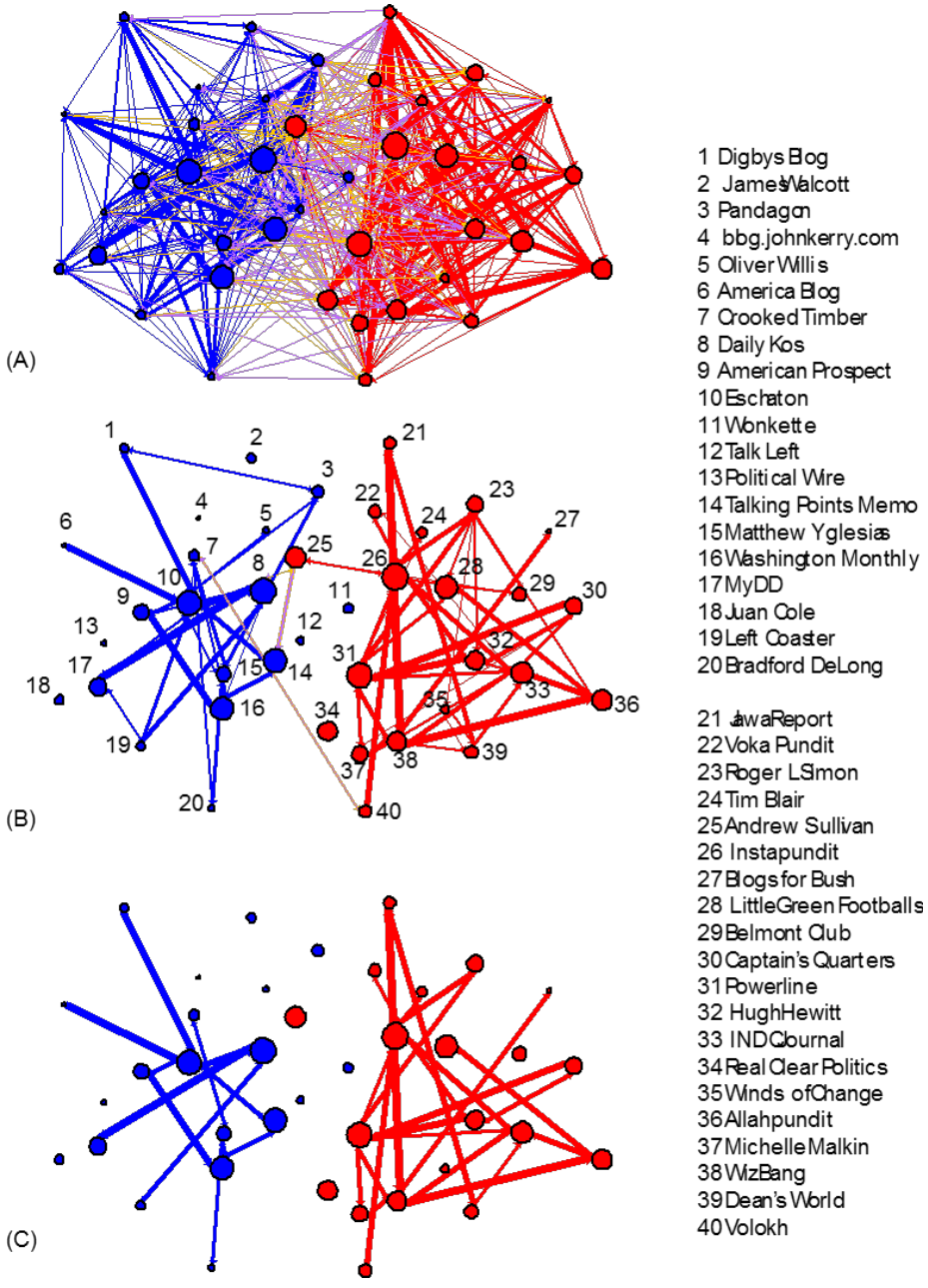


Figure 17: Ejemplo.

Newman y Girvan. Finding and evaluating community structure in networks. Phys Rev E 69:026113 (2004)

El propósito de los dos enfoques es descubrir grupos de nodos relacionados en la red y, si es posible, la estructura jerárquica correspondiente, a partir de la información proporcionada por la topología de la red

- Una de las heurísticas más extendidas es eliminar iterativamente los puentes entre grupos de nodos (Girvan y Newman, 2002)
- Estos métodos devuelven particiones disjuntas del conjunto de nodos, cada nodo pertenece a una única comunidad (no permiten el solapamiento de comunidades).
- Existen algunos que consideran dicho solapamiento, como el clique percolation method de Palla y otros (2005). Son especialmente útiles en Ciencias Sociales

Fuentes: Girvan y Newman. Community structure in social and biological networks.Proc Natl Acad Sci USA 99:7821–7826 (2002). Palla y otros. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435:814–818 (2005)

5.5 Medida de modularidad

La modularidad Q es una función de calidad que mide la calidad de una partición concreta de una red de comunidades. Se define como la diferencia entre el número de enlaces existentes entre los grupos y el número de enlaces esperado en una red aleatoria equivalente.

$$Q = /frac12L \sum_{ij} [A_{ij} - \frac{k_i k_j}{2L}] \delta(c_i, c_j)$$

donde

- L es el número de enlaces de la red.
- A_{ij} es la matriz de adyacencia.
- La probabilidad de un enlace entre 2 nodos es proporcional a sus grados. $\frac{k_i k_j}{2L}$
- $\delta(c_i, c_j)$ vale 1 si los nodos son de la misma comunidad.
- $Q \epsilon [-1, 1]$

Fuente: Clauset, Newman, Moore. Finding community structure in very large networks. Phys Rev E 70: 066111 (2004)

La idea básica es que una red muestra una estructura modular coherente si el número de enlaces entre comunidades es menor que el esperado en una red aleatoria.

Cuanto mayor es el valor de Q, mejor es la partición, es decir, las comunidades encontradas están densamente conectadas internamente (hay más enlaces de los que cabría esperar aleatoriamente) y dispersamente conectadas entre sí.

En una red aleatoria, Q=0. En la práctica, una modularidad de 0.3 es un buen valor.

Esta medida se usa tanto para comparar la calidad de distintas particiones como para diseñar métodos de descubrimiento de comunidades que traten de maximizar su valor.

5.6 Referencias

Clauset, Newman, Moore. Finding community structure in very large networks. Phys Rev E 70: 066111 (2004)

Zachary. An information flow model for conflict and fission in small groups. J. Anthropol. Res. 33: 452-473 (1977)

Exploratory Social Network Analysis with Pajek

Rosvall y Bergstrom. Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. USA 105: 1118-1123 (2008)

Girvan y Newman. Community structure in social and biological networks.Proc Natl Acad Sci USA 99:7821–7826 (2002).

Palla y otros. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435:814–818 (2005)

Wasserman y Faust. Social Network Analysis. Cambridge University Press; 1994