Minería de Medios Sociales
Máster en Ciencia de Datos e Ingeniería de Computadores
Bloque I: Redes Sociales y Minería de Datos en Redes

Sesión I.2: Análisis de Redes Sociales

Oscar Cordón García
Dpto. Ciencias de la Computación e Inteligencia Artificial. Universidad de Granada
ocordon@decsai.ugr.es

ANÁLISIS DE REDES SOCIALES

HISTORIA

Orígenes del Análisis de Redes Sociales

El **Análisis de Redes Sociales** (*Social Network Analysis*, SNA) se centra en el descubrimiento de patrones de interacción entre **actores sociales** en redes sociales

Es un área de investigación metodológica interdisciplinar con contribuciones de la **Sociología**, la **Psicología Social**, la **Antropología**, la Física, las Matemáticas y la Informática, entre otras

Los orígenes del SNA, como una base para el desarrollo de conceptos sociológicos útiles, puede fecharse a **comienzos de la década de 1930**, cuando Moreno desarrolló el **enfoque sociométrico** como una forma de conceptualizar la estructura de las relaciones sociales establecidas entre pequeños grupos de individuos

Estos lazos interpersonales entre miembros de un grupo fueron representados mediante los denominados **sociogramas**, que pueden definirse como gráficos en los que los individuos se representan como nodos y las relaciones entre ellos como líneas

Esos diagramas resultaron ser muy útiles para descubrir las estructuras ocultas de los grupos mediante la identificación de protagonistas, alianzas y subgrupos, entre otras cosas

Moreno JL. Who Shall Survive? New York: Beacon House; 1953

ANÁLISIS DE REDES SOCIALES (*SOCIAL NETWORK ANALYSIS*)

El objetivo principal del **Análisis de Redes Sociales** (SNA) es examinar tanto los contenidos como los **patrones de relación en redes sociales para entender las relaciones entre los actores** y las implicaciones de esas relaciones

Es un área interdisciplinar entre las ciencias sociales, la estadística, la teoría de grafos, la complejidad y la informática

Son tareas habituales del SNA:

- identificar los actores más influyentes, prestigiosos o centrales de la red, mediante medidas estadísticas,
- identificar *hubs* y autoridades, usando algoritmos de análisis de enlaces,
- determinar patrones de interacción comunes entre actores, mediante medidas de niveles de interacción, y
- descubrir grupos de actores cohesionados, con técnicas de detección de comunidades

Wasserman S., Faust, K. Social Network Analysis. Methods and Applications. Cambridge University Press; 2008

EJEMPLOS DE APLICACIÓN DEL ANÁLISIS DE REDES SOCIALES

- **Marketing viral: maximización de la difusión "boca a boca" de productos** de una compañía dirigiéndose a los clientes de mayor valor en la red (aquellos con una mayor influencia y soporte)

- Análisis de las redes de llamadas telefónicas en compañías de telecomunicaciones para **identificación de perfiles de los usuarios y recomendación de tarifas personalizadas** de acuerdo a dichos perfiles

- Uso para **predicción de la deserción de clientes** (*churn prediction*) identificando cambios en sus patrones de contactos telefónicos

- **Detección de fraudes**, por ejemplo, en comunicaciones organizacionales (conjunto de datos de Enron) para analizar la frecuencia y la dirección de los envíos de e-mails formales/informales que pueden revelar los patrones de comunicación entre empleados y jefes
  Estos patrones pueden ayudar a identificar personas implicadas en actividades fraudulentas

TIPOS DE MEDIDAS DE ANÁLISIS DE REDES SOCIALES

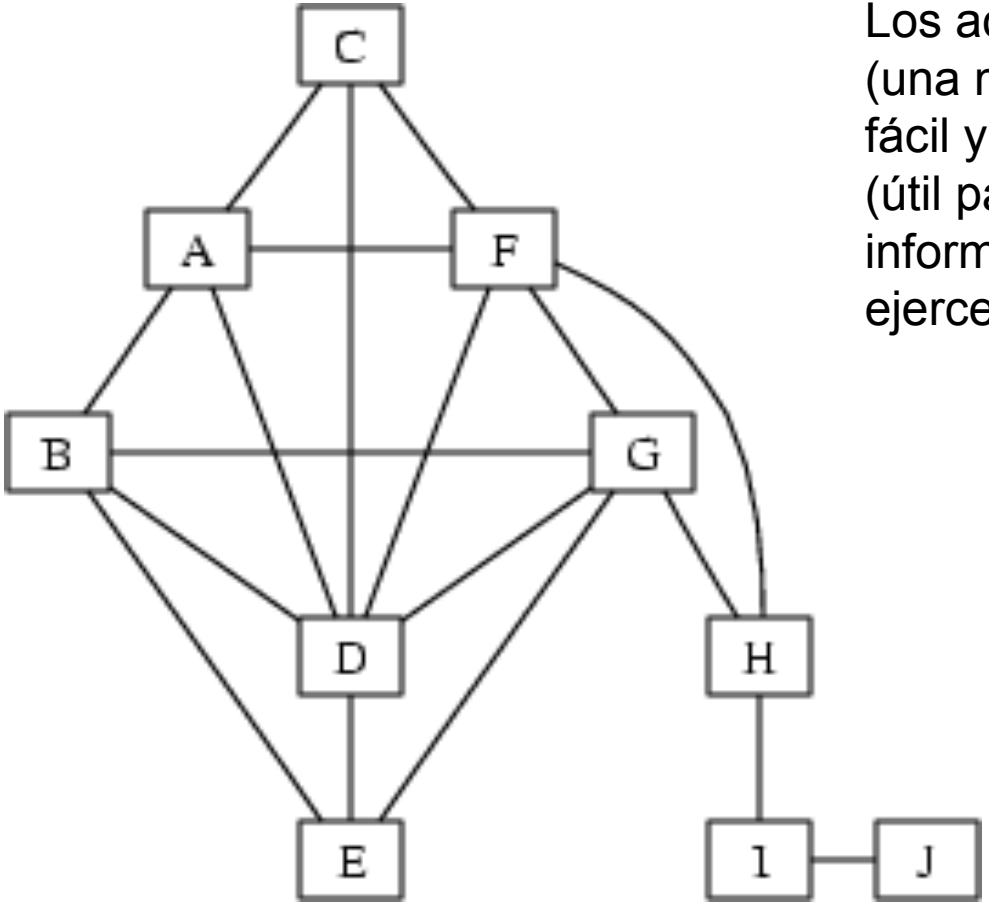
De la misma forma que en el análisis de redes en general, el análisis de la estructura de las redes sociales pretende entender el comportamiento de los sistemas complejos (en este caso, *sistemas sociales*) que generan dichas redes

Existen dos tipos de medidas:

- **Medidas locales (a nivel de actores)**: Todas ellas están basadas en el concepto general de **centralidad** (redes no dirigidas) o **prestigio** (redes dirigidas), una medida general de la posición de un actor en la estructura global de la red social

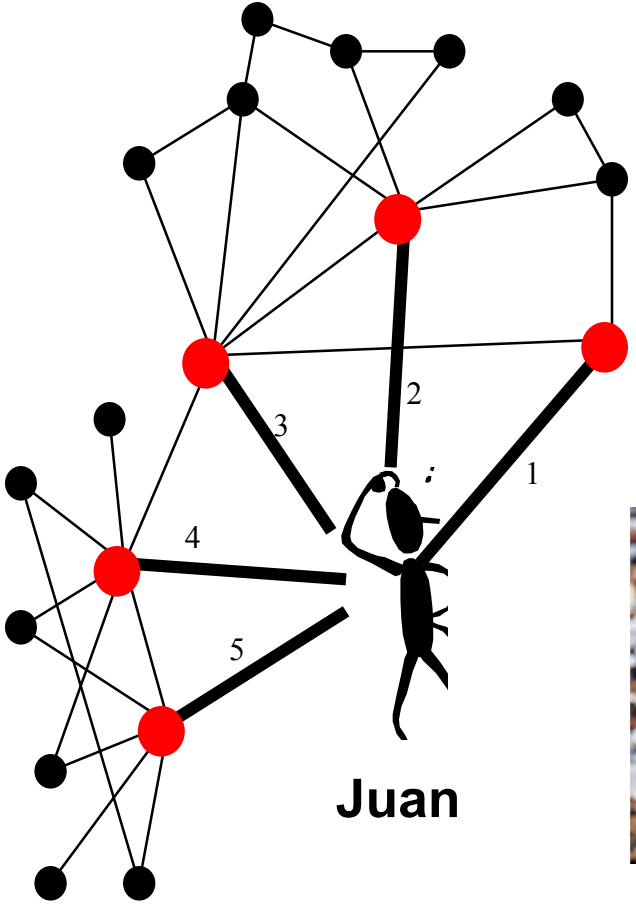
  Se usan para identificar los ***actores clave*** de la red. Muestran como las relaciones se concentran en unos pocos individuos, dando una idea de su *poder social*

- **Medidas globales (a nivel de red)**: Proporcionan información más compacta que permite evaluar la estructura global de la red, aportando información sobre propiedades importantes de los fenómenos sociales subyacentes

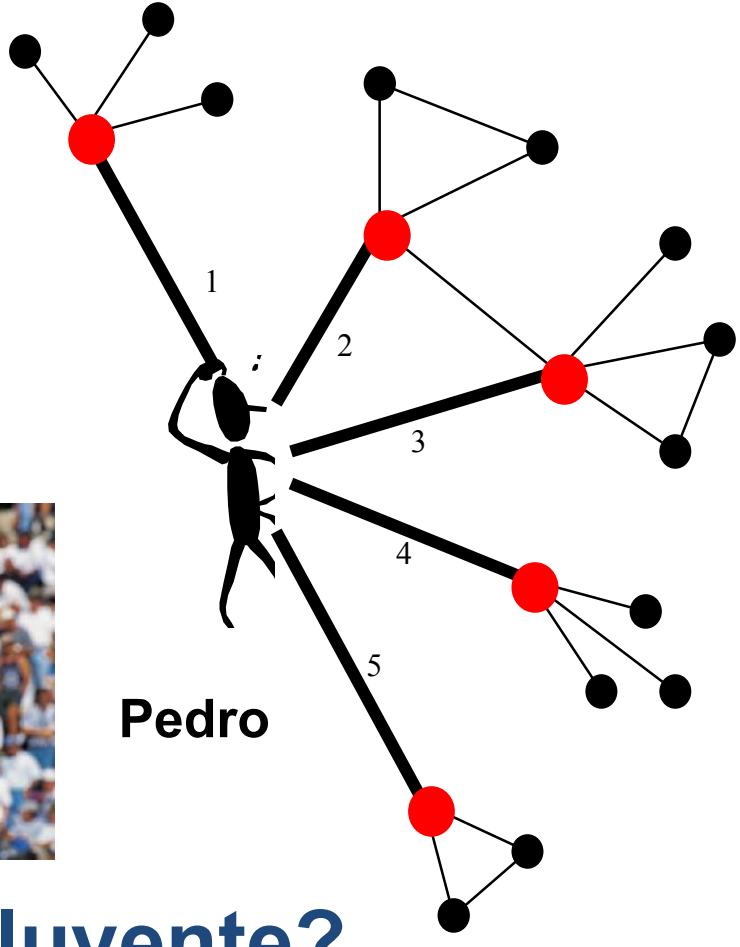


Los actores con una **"posición más central"** (una mayor centralidad) tienen un acceso más fácil y rápido a los demás actores de la red (útil para acceder a recursos como información) y una mayor capacidad para ejercer un control del flujo entre ellos

¿Quién es central en esta red?



Juan



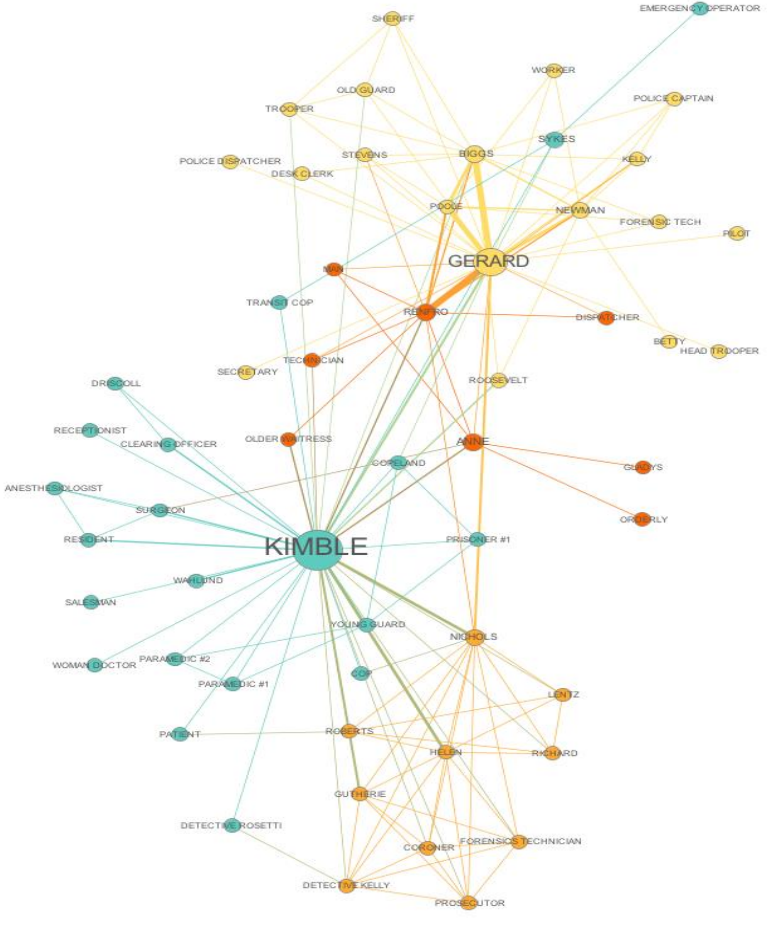
Pedro

¿Quién es más influyente?

MEDIDAS LOCALES DE CENTRALIDAD

Ejemplo de Centralidad

The Fugitive (1993)



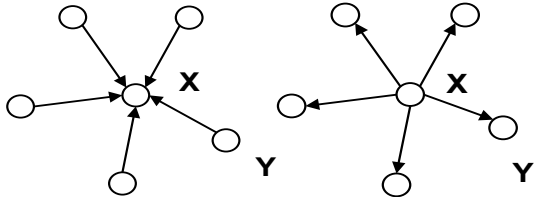
http://es.wikipedia.org/wiki/El_fugitivo_película_de_1993

moviegalaxies.com

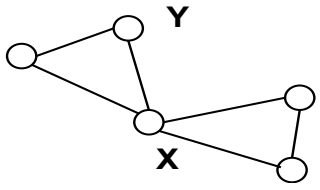
MEDIDAS LOCALES DE CENTRALIDAD

Existen varias medidas distintas de centralidad:

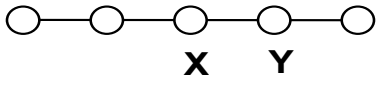
1. **Grado**:



2. **Intermediación** (*betweenness*):



3. **Cercanía** (*closeness*):



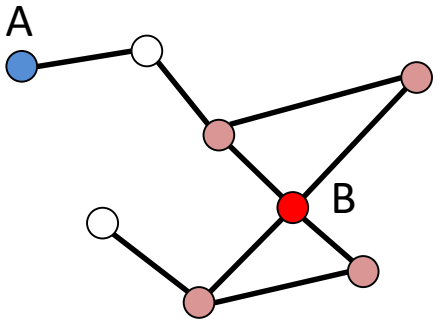
4. **Excentricidad**

5. **Centralidad de vector propio**

Es conveniente conocer bien las características de cada medida y usar varias

P.ej., las Centralidades de grado son medidas importantes pero no tienen en cuenta la estructura global de la red

No dirigida



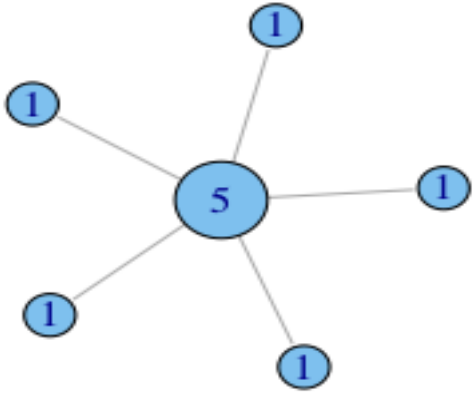
Centralidad de grado de un actor (C_D): número de enlaces que lo conectan con otros

$$C_D(\mathrm{A}) = k_{\mathrm{A}} = 1 \qquad C_D(\mathrm{B}) = k_{\mathrm{B}} = 4$$

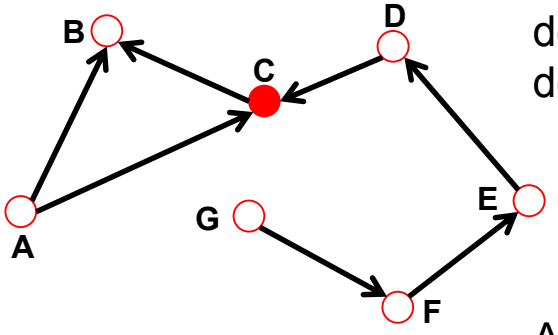
C_D(i) se define en {0,*g*-1}, siendo *g* el número de nodos de la componente conexa

Interpretación: ***Los actores con más amigos son más centrales***

Sólo mide la importancia con respecto a los vecinos más cercanos. Se asume que *las conexiones de los amigos no importan*, sólo importa lo que ellos pueden hacer directamente (p.ej. “ir a tomarse una cerveza contigo”, “ayudarte a hacer una práctica”, etc.)



Dirigida



En redes dirigidas, se define el **Prestigio** de entrada (*in-degree*), denominado **Soporte**, y el **Prestigio** de salida (*out-degree*), denominado **Influencia**:

$$P_D^{in}(C) = k_C^{in} = 2 \qquad P_D^{out}(C) = k_C^{out} = 1$$

Ambos se definen en {0,*g*-1}

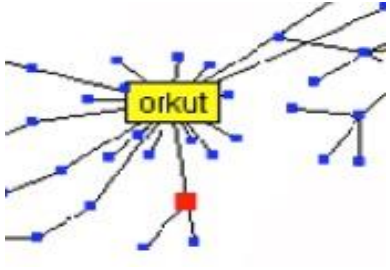
Interpretación Soporte: ***Los actores que reciben muchos enlaces son <u>prominentes</u>***
La idea básica es que muchos actores procuran tener enlaces directos a ellos, por lo que se puede considerar como una medida de importancia

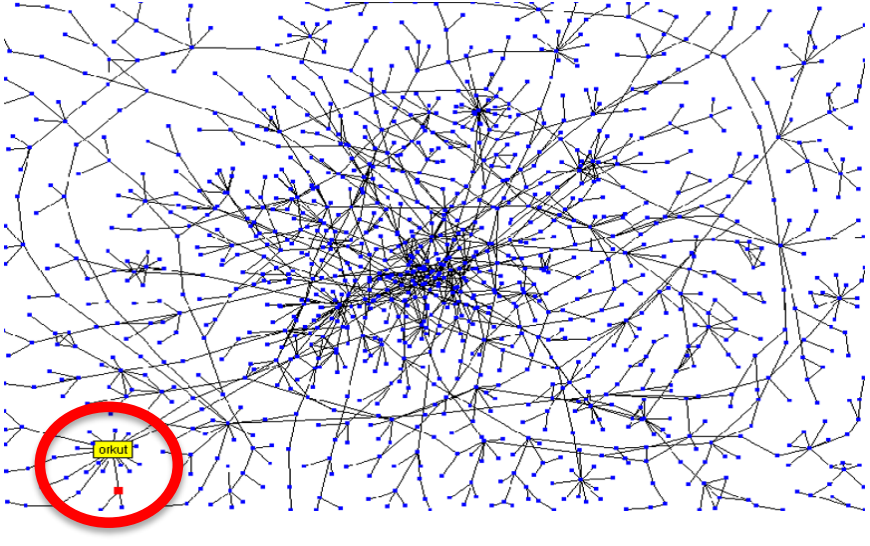
Interpretación Influencia: ***Los actores que tienen muchas conexiones directas con otros son <u>influyentes</u>***
Se entiende que pueden intercambiar o transferir información rápidamente a muchos otros (*argumento de la fortaleza de las conexiones débiles*)

MEDIDAS LOCALES DE CENTRALIDAD

Comportamiento

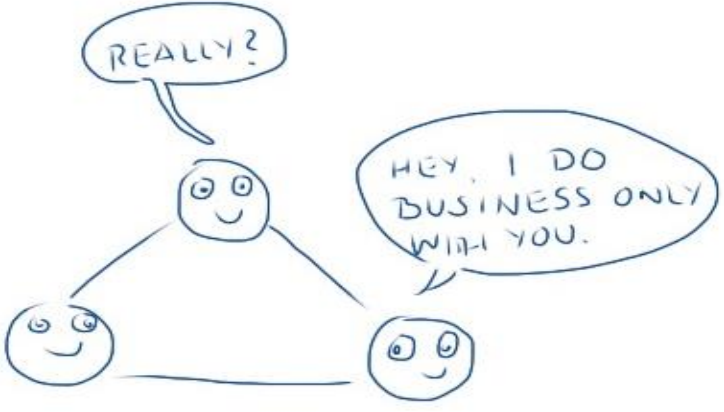
Stanford Social Web (ca. 1999)

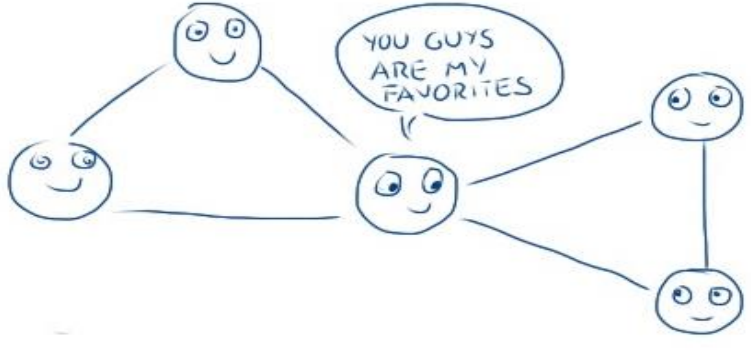




Red de páginas web personales en Stanford

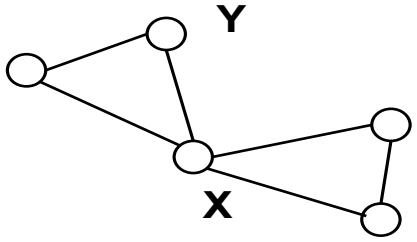
El grado no captura las "correduría" (*brokerage*)





La **intermediación** es una medida pensada para capturar la correduría:

$$C_B(i) = \sum_{j,k \in V(G)/v} g_{jk}(i) / g_{jk}$$



donde g_{jk} es el número de caminos mínimos que conectan cualquier par de nodos *j* y *k* (normalmente 1) y $g_{jk}(i)$ es el número de esos caminos que incluyen al actor *i*. C_B(i) se define en {0, (g-1)·(g-2)} en redes dirigidas y en {0, (g-1)·(g-2)/2} en no dirigidas

Intuición: Ver al actor con una posición más favorable en la medida en que dicho actor esté situado entre los caminos geodésicos de todos los demás

En otras palabras, cuantos mas nodos que necesiten pasar por mi para hacer sus conexiones indirectas por los caminos más cortos, más central seré yo

Es habitual considerar la medida normalizada (redes no dirigidas):

$$C'_B(i) = \frac{C_B(i)}{(g-1)(g-2)/2}$$

Número de pares de actores excluyendo el propio nodo i

Los actores con una intermediación alta ocupan roles críticos en la estructura de la red puesto que suelen ocupar una posición que les permite trabajar como **interfaces entre subgrupos de actores fuertemente unidos**

Son elementos vitales en la conexión entre distintas regiones de la red

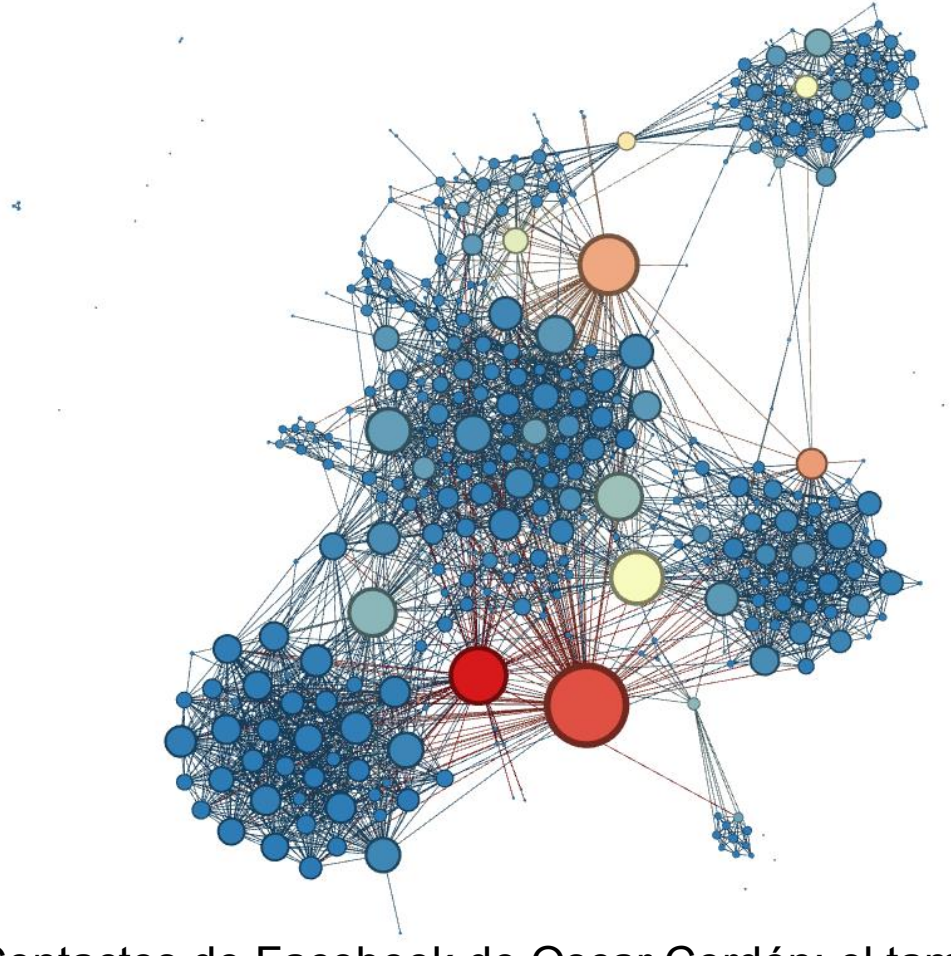
En la perspectiva de las redes sociales, **las interacciones entre dos actores no adyacentes pueden depender de otros actores del conjunto, especialmente de aquellos situados en los caminos entre ambos**

Estos actores se denominan también **porteros** (*gatekeepers*) porque tienden a controlar el flujo de información entre comunidades

La intermediación también puede calcularse para las relaciones, midiendo el grado en que hacen posible otras conexiones. Determina **puentes locales**, enlaces con intermediaciones altas: $$C_B^{'}(e) = \sum_{j,k \in V(G)} g_{jk}(e) / g_{jk}$$

MEDIDAS LOCALES DE CENTRALIDAD

Grado vs. Intermediación

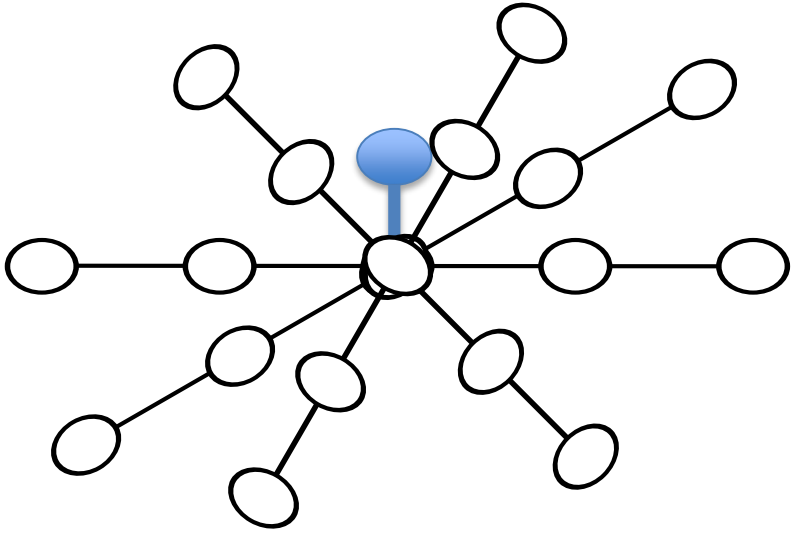


Red Personal de Contactos de Facebook de Oscar Cordón: el tamaño de los nodos indica el **grado** y el color la **intermediación** (más azul, menor valor; más rojo, mayor valor)

MEDIDAS LOCALES DE CENTRALIDAD | Cercanía (1)

La **cercanía** es una tercera forma alternativa de medir la centralidad que se plantea el hecho de que puede no ser tan importante tener muchos amigos directos ni estar situado “entre” otros actores

En este caso, **se le da importancia a “estar en medio de las cosas”, no demasiado lejos del centro**, para lo cual no es necesario estar en una posición de correduría

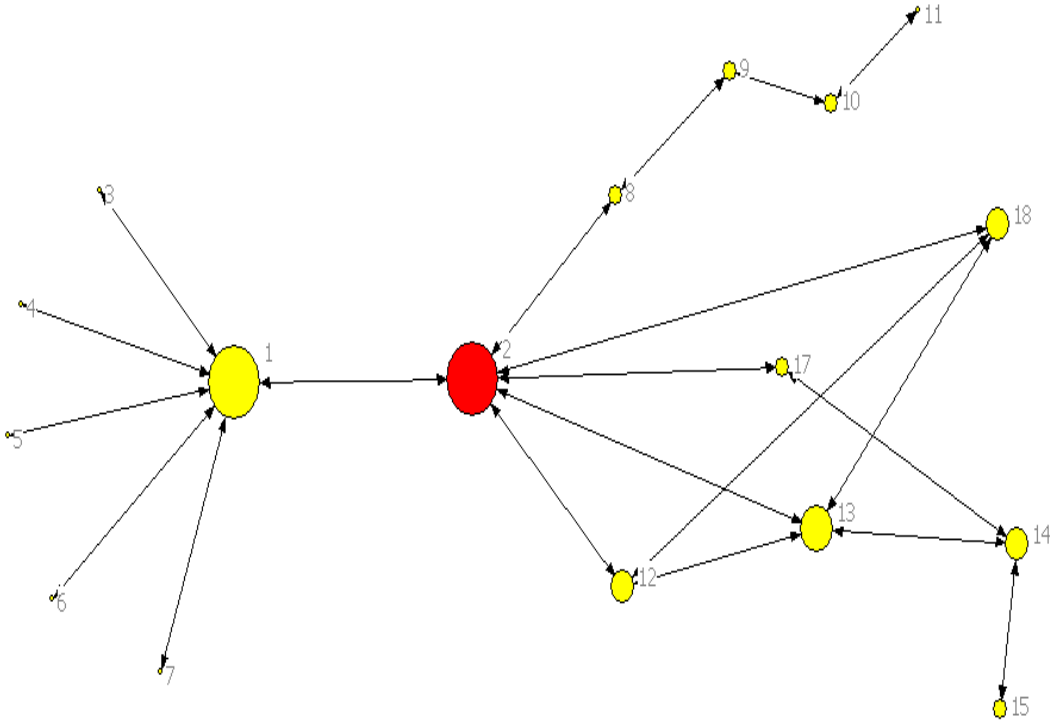


Se enfatiza la distancia de un actor a otros en la red al concentrarse en la distancia geodésica de cada actor con todos los demás

La suma de estas distancias geodésicas (distancias de los caminos mínimos) para cada actor es la lejanía de dicho actor al resto. **La inversa de dicha suma es la medida de cercanía**

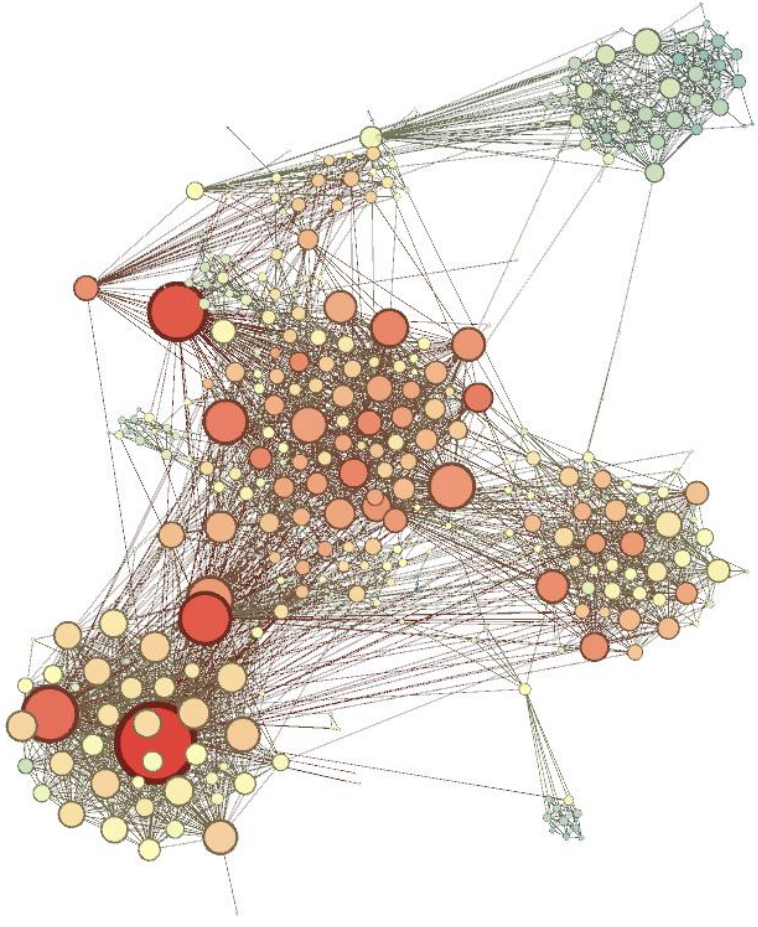
Fórmulas de la **Centralidad de cercanía** (sin normalizar, C_C, y normalizada, C'_C):

$$C_C(i) = \frac{1}{\sum_{g} d(i,j)} \qquad C'_C(i) = \frac{C_C(i)}{\frac{1}{g-1}} = (g-1)\cdot C_C(i)$$



Actor	Lejanía	CercaníaN
2	34.000	50.000
1	40.000	42.500
13	42.000	40.476
17	44.000	38.636
8	44.000	38.636
12	45.000	37.778
18	45.000	37.778
14	52.000	32.692
6	56.000	30.357
5	56.000	30.357
7	56.000	30.357
3	56.000	30.357
4	56.000	30.357
9	56.000	30.357
15	66.000	25.758
10	70.000	24.286
16	82.000	20.732
11	86.000	19.767

MEDIDAS LOCALES DE CENTRALIDAD | Grado vs. Cercanía



Red Personal de Contactos de Facebook de Oscar Cordón: el tamaño de los nodos indica el **grado** y el color la **cercanía** (más azul, menor valor; más rojo, mayor valor)

Otra medida local de centralidad basada en distancias es la **Centralidad de excentricidad (C_E)**. Se define como la inversa de la **excentricidad** (la máxima distancia geodésica) entre un actor y cualquier otro actor de la red:

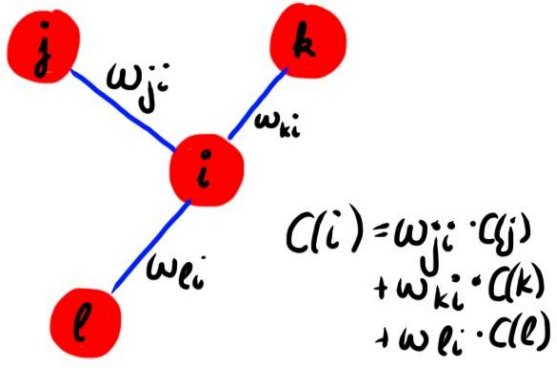
$$C_E(i) = \frac{1}{\max\limits_{j \in V(G)/i} d(i,j)}$$

$$C_E^{'}(i) = C_E(i) / g - 1$$

Los actores con un mayor valor de excentricidad se denominan **actores periféricos**, los de menor valor forman el **centro de la red**

La **Centralidad de vector propio** se basa en que la centralidad de un nodo concreto depende de cómo de centrales sean sus vecinos (**prominencia**)

La idea básica es que el poder y el status de un actor (***ego***) se define recursivamente a partir del poder y el status de sus vecinos (***alters***)



w_{ij} (a_{ij}) corresponde a la entrada de la matriz de adyacencia. Puede ser binaria {0,1} o un peso numérico

La medida es válida para redes dirigidas (**Prestigio de rango**) y no dirigidas

Es una versión más elaborada de la Centralidad de grado al asumir que no todas las conexiones tienen la misma importancia. No se tiene en cuenta la cantidad sino la calidad de las mismas

Centralidad de vector propio (2)

La medida de Centralidad de vector propio, C_{VP}, se define como una combinación lineal (o una **suma**, si los enlaces no están ponderados) de los valores de todos los actores que apunten a *i*:

$$C_{VP}(i) = a_{1i} \cdot C_{VP}(1) + a_{2i} \cdot C_{VP}(2) + \ldots + a_{ni} \cdot C_{VP}(n)$$

Para calcular los valores de C_{VP} para los *n* actores se construye un sistema de *n* ecuaciones con *n* incógnitas que se representa de forma matricial

Si $\boldsymbol{C}=(C_{VP}(1), \ldots, C_{VP}(n))^T$ es el vector transpuesto que almacena los *n* valores de C_{VP} (***C*** es un vector columna) y ***A*** es la matriz de adyacencia, entonces:

$$\boldsymbol{C} = \boldsymbol{A}^T \cdot \boldsymbol{C}$$

Esta ecuación coincide con la **ecuación característica para encontrar los vectores y valores propios de la matriz** $\boldsymbol{A}^T$. ***C*** es un vector propio de $\boldsymbol{A}^T$

Es habitual considerar la medida normalizada:

$$c_i = \frac{1}{\lambda}\sum_{j=1}^{n} a_{ij}\cdot c_j \qquad \boldsymbol{C} = \frac{1}{\lambda}\boldsymbol{A}^T\cdot\boldsymbol{C} \qquad \lambda\cdot\boldsymbol{C} = \boldsymbol{A}^T\cdot\boldsymbol{C}$$

donde λ es una constante que equivale al mayor valor absoluto del vector propio dominante de ***A***

Existen distintos algoritmos para calcular ese vector propio ***C***, como el **Método de las Potencias**, usado por Google en el Pagerank (redes dirigidas)

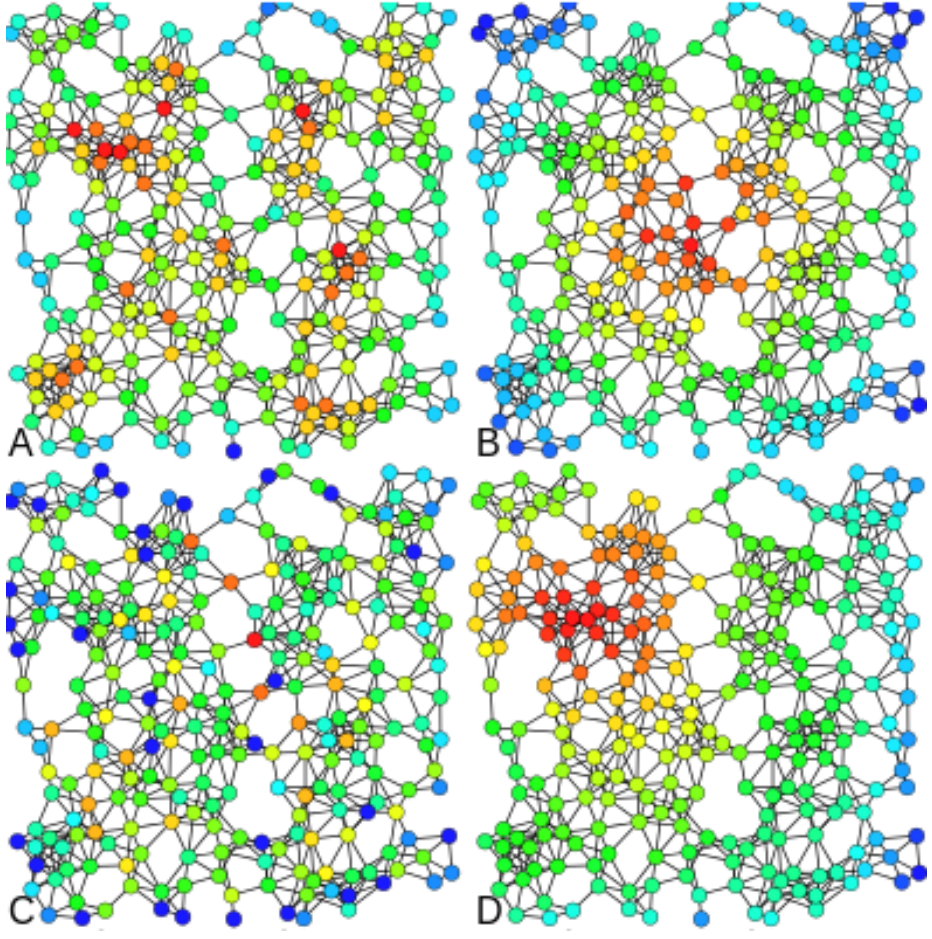
http://es.wikipedia.org/wiki/Método_de_las_potencias

Como la fórmula es circular, el Método de las Potencias es iterativo. Sin embargo, se deben cumplir unas condiciones específicas para poder aplicarlo

MEDIDAS LOCALES DE CENTRALIDAD | Comparativa

A) centralidad de grado; B) cercanía; C) intermediación; D) centralidad de vector propio

azul = menor valor



rojo = mayor valor

MEDIDAS GLOBALES

Existen varias medidas globales en SNA. La mayoría son las mismas empleadas para analizar cualquier otro tipo de red:

1. **Densidad**
2. **Diámetro y Radio** (longitud del máximo y mínimo camino geodésico)
3. **Distancia media**
4. **Grado Medio**

Existen medidas adicionales para **analizar patrones de interacción** como el **Coeficiente Global (Medio) de Clustering**, que mide la transitividad, y la **Reciprocidad**

ALGUNAS APLICACIONES DEL ANÁLISIS DE REDES SOCIALES

APLICACIONES | Análisis de Influencia de los Líderes Chinos (1)

http://www.iaventures.com/what-is-the-question-that-sells-the-business

Estudio realizado por Drew Conway, científico de datos:

Collaboration with Recorded Future, a Boston start-up that specializes in longitudinal entity extraction from the massive amount of open-source data generated daily. The question was:

Who are the most central members of the China's leadership as we enter 2012?

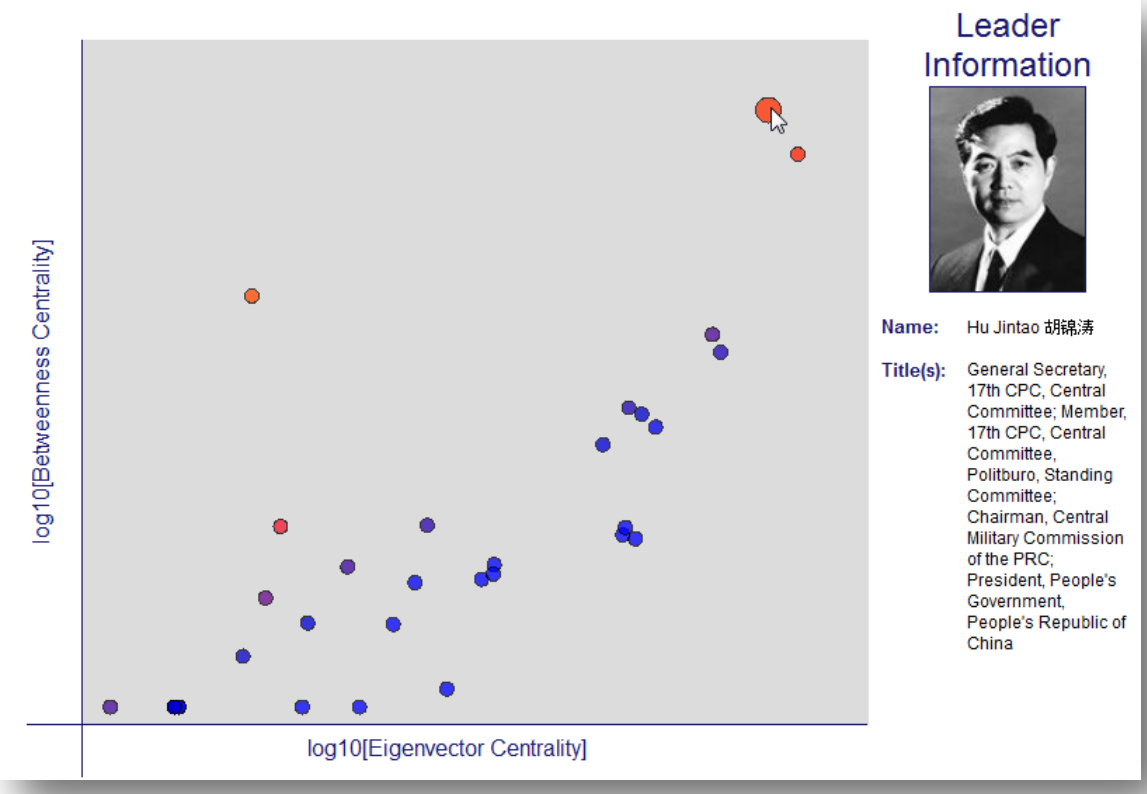
For our analysis we focused on the China's leadership, as defined by the CIA World Factbook, and extracted all of named entities in their data for 2011 (over 4 billion events) for which any of the 33 official Chinese leaders appear

The result is a dataset with over 150,000 entities; including people, organizations, and places

To answer our questions, however, I used the **co-occurrence** of these **entities** in sentence fragments to build a large **network** of these entities

APLICACIONES | Análisis de Influencia de los Líderes Chinos (2)

To visualize the results a **simple interactive scatter-plot of centrality measures** was used:



This is a great example of how an analysis can highlight the power of big data, lead to better informed follow-up questions, and convey the value of the business providing it

APLICACIONES

Análisis de Cuentas de Twitter sobre “Information Visualization” (1)

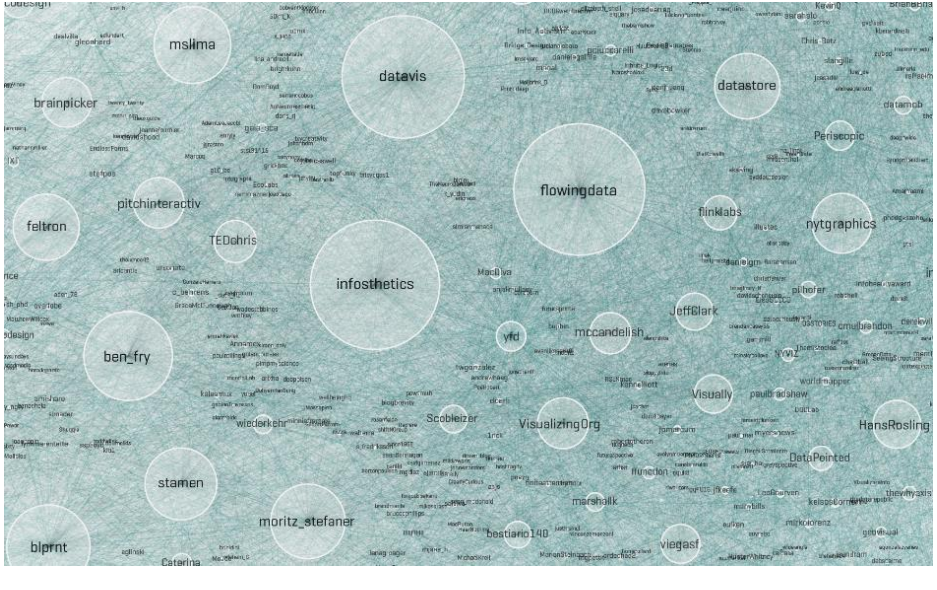
http://well-formed-data.net/archives/642/the-vizosphere

http://blogger.ghostweather.com/2011/09/combing-through-infovis-twitter-network.html

Mapa de 1645 cuentas de Twitter sobre “*information visualization*” por Moritz Stefaner

Se seleccionaron unas cuentas *semilla* y se extrajeron los seguidores y amigos mediante el API de Twitter. **Se filtraron aquellas cuentas con menos de 5 enlaces a las cuentas originales**

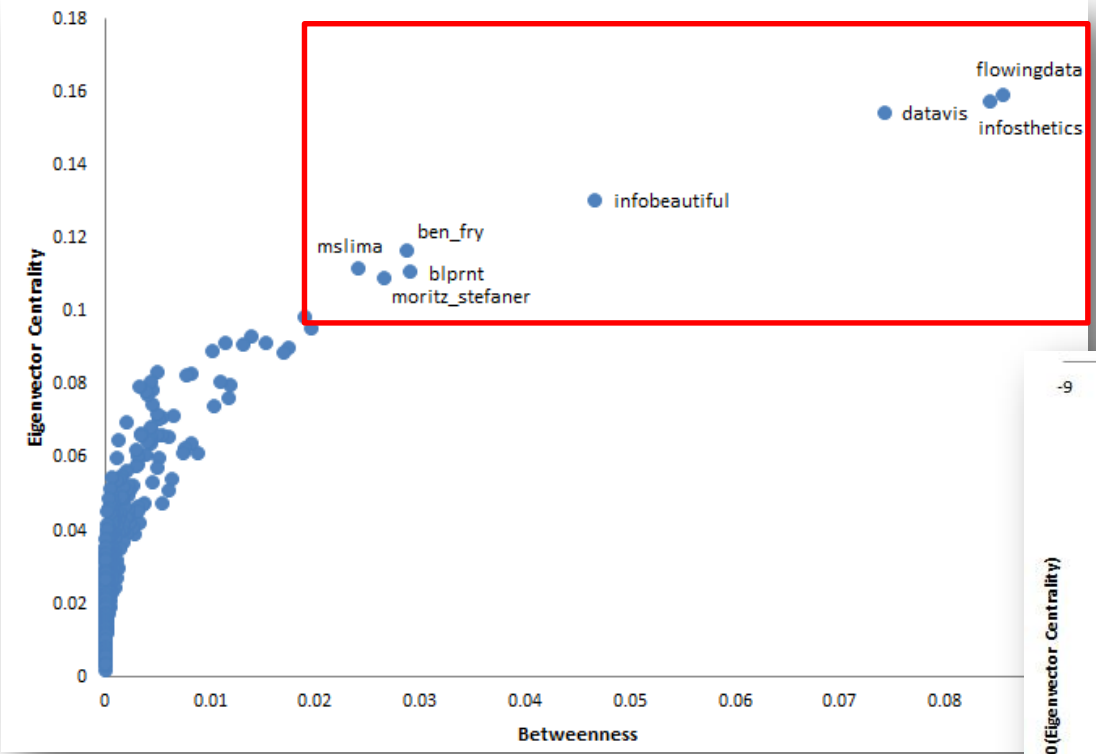
El tamaño de los nodos indica el número de seguidores en esta red concreta

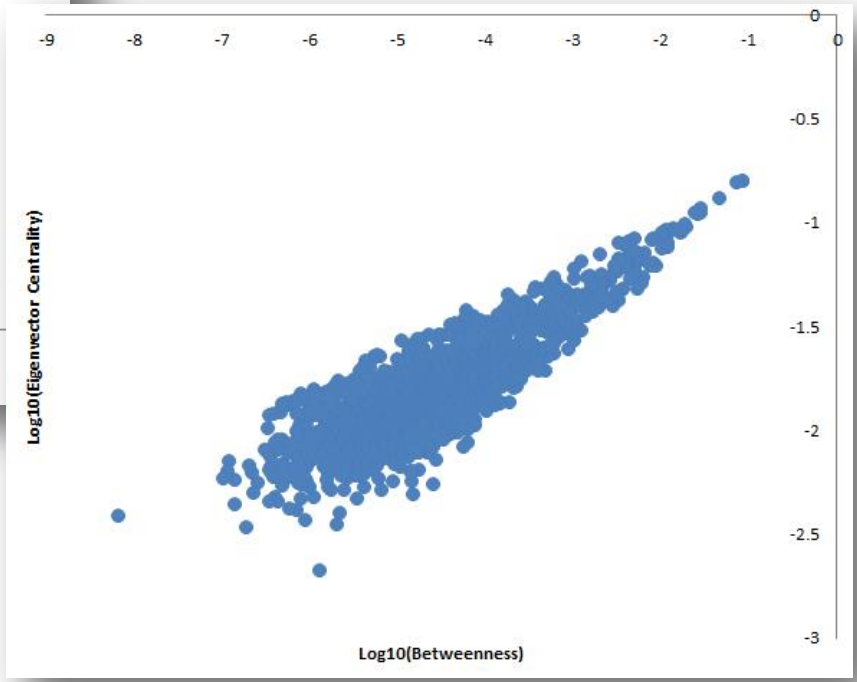


Se emplearon correlaciones entre medidas de centralidad para **determinar la influencia de las cuentas**

APLICACIONES

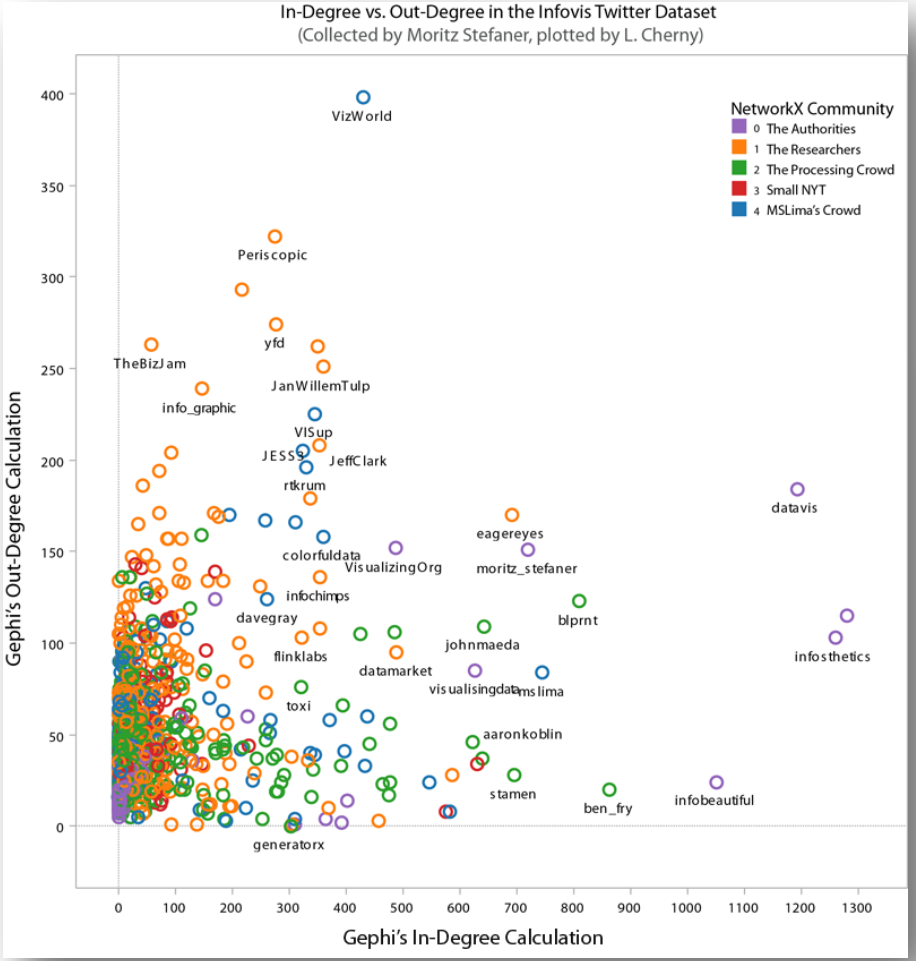
Análisis de Cuentas de Twitter sobre “Information Visualization” (2)



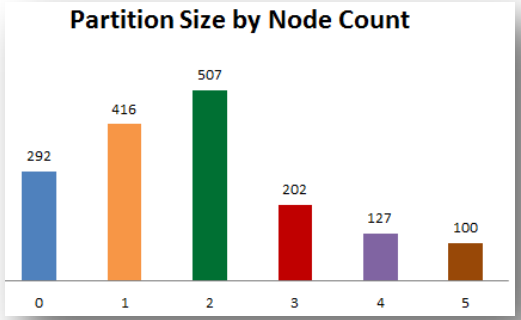


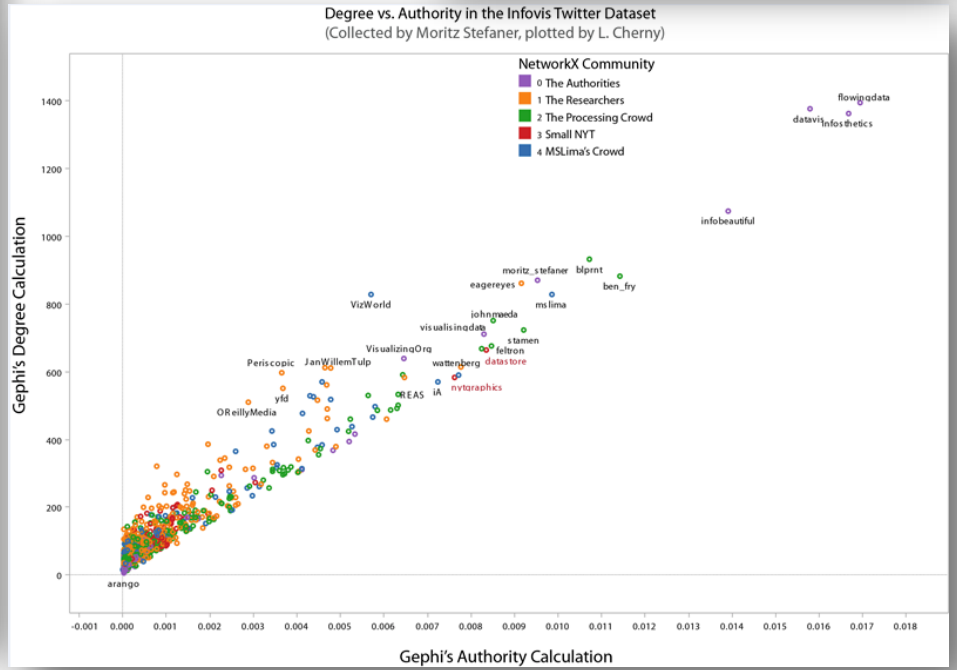
APLICACIONES

Análisis de Cuentas de Twitter sobre “Information Visualization” (3)



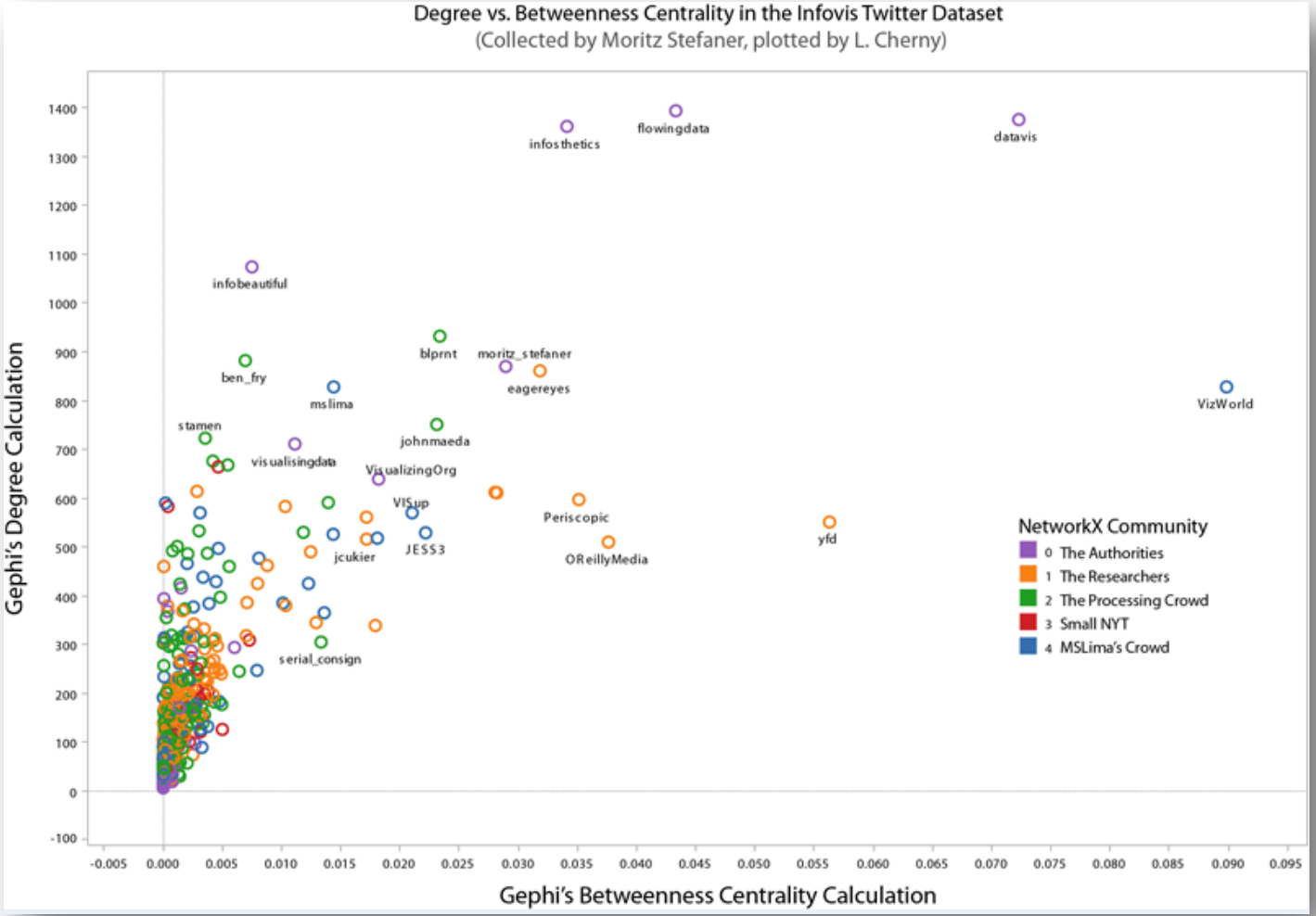
Detección de comunidades con el método de Lovaina





APLICACIONES

Análisis de Cuentas de Twitter sobre “Information Visualization” (4)



Referencias y Agradecimientos

Para elaborar las transparencias de este curso, he hecho uso de algunos materiales desarrollados por expertos en el área disponible en Internet:

- "Network Science Interactive Book Project" del Laszlo Barabasi Lab. Northeastern University: http://barabasilab.com/networksciencebook
- K. Lerman. "Social Media. A Responsible User's Guide". University of South California: http://www.isi.edu/integration/people/lerman/talks.html
- Curso on-line "Social Network Analysis" de Lada Adamic, Coursera, Universidad de Michigan: https://www.coursera.org/course/sna
- L. Cherny. "Simplifying Social Network Visualizations". Ghostweather Research & Design, LLC: http://es.slideshare.net/arnicas/simplifying-social-network-diagrams?related=1