

Análisis de redes sociales

¿Cómo identificar personas significativos en una red social y analizar su papel?

¿Qué características globales de la estructura de las redes interesa medir?

¿Es posible explicar analíticamente la estructura y formación de redes sociales?

Análisis de redes sociales

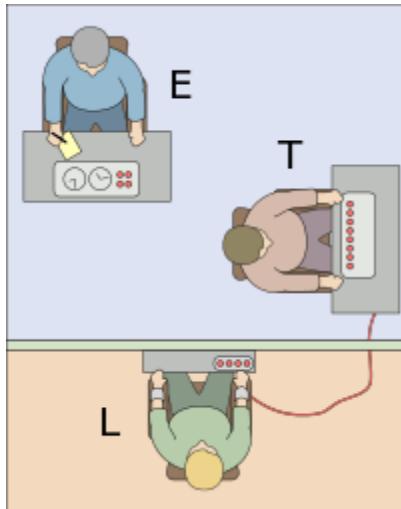
1. Introducción
2. Topologías de red: métricas y estadísticas
3. Subdivisión de redes
4. Modelos de formación de redes
5. Procesos de difusión

1. Introducción

“The” Milgram experiment (1963)



Stanley Milgram
(1933-1984)



¿Estamos todos conectados?

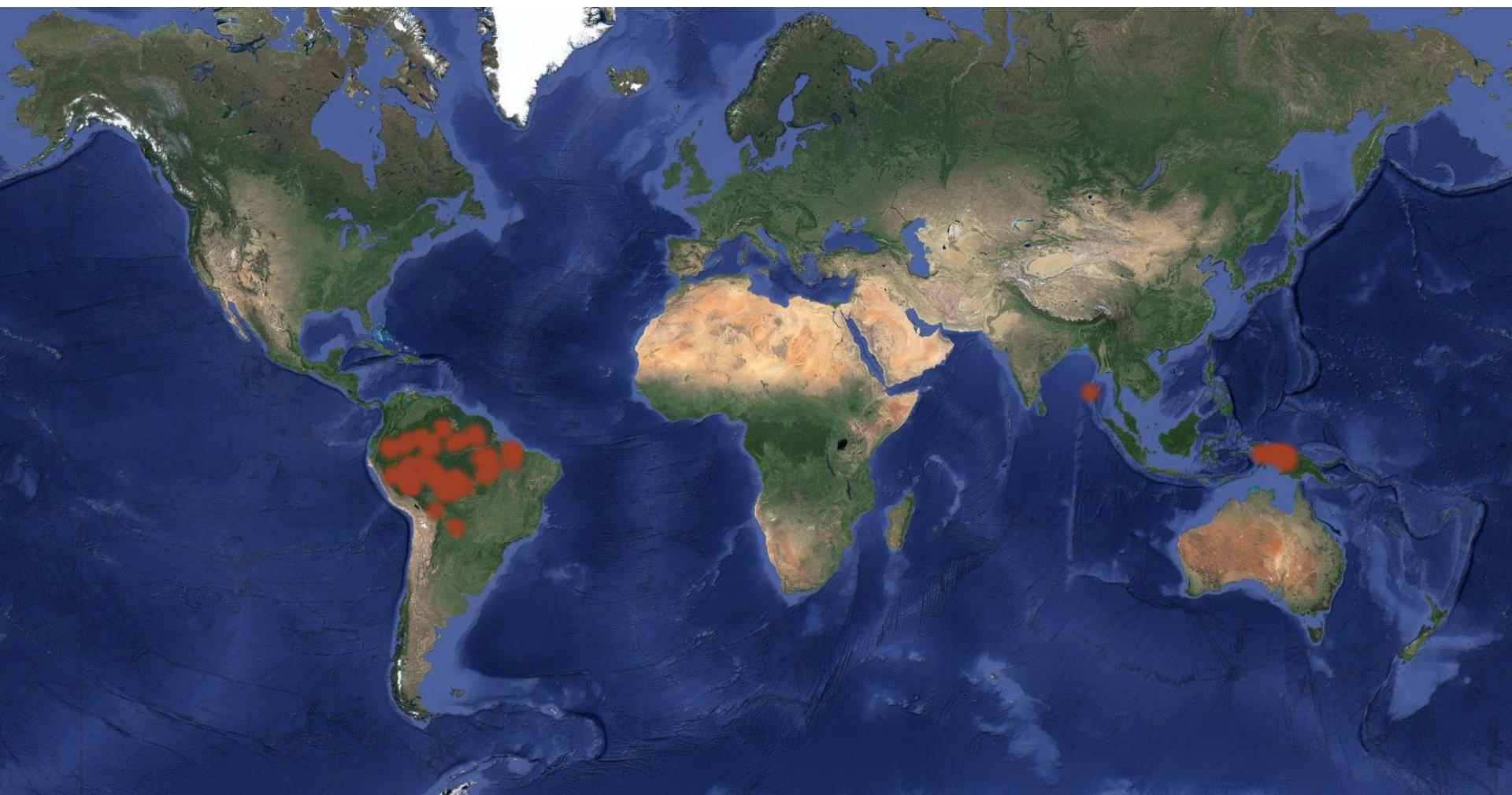
Sentinel Island



Acre, Brasil

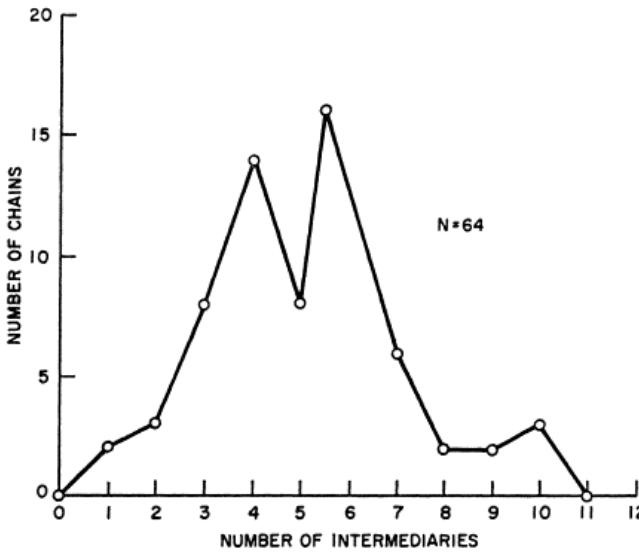


¿Estamos todos conectados?



Experimento small-world Milgram (1967)

¿Cuál es la distancia promedio entre dos personas al azar?

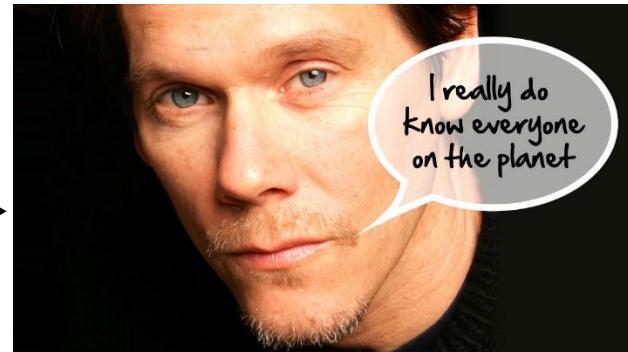


- ◆ Varias personas origen elegidas “al azar”, una persona destino
- ◆ Tarea: hacer llegar un paquete a la persona destino por medio de una cadena de amigos
 - Cada persona añade su nombre en una lista + notifica a Milgram
- ◆ 64/296 cadenas llegaron al destino
- ◆ Resultado: distancia promedio ~ 6.2 (nº arcos)

“Six degrees of separation” en la cultura popular



Erdős nr = 4.65 promedio
entre matemáticos en activo



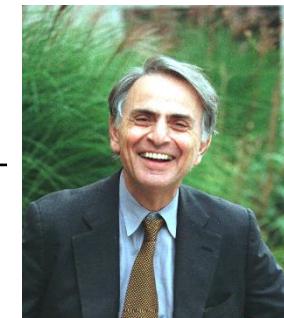
Bacon nr, avg ~ 3, máx = 10
oracleofbacon.org

4



2

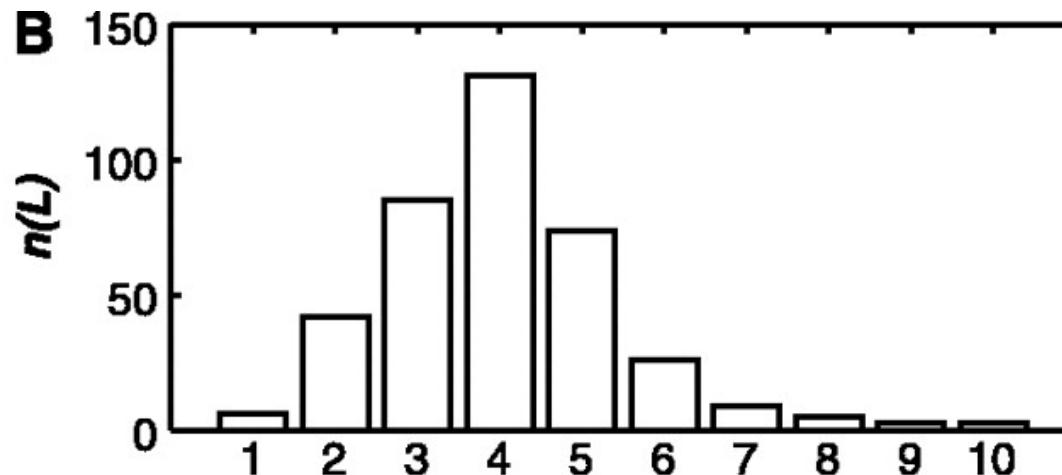
4



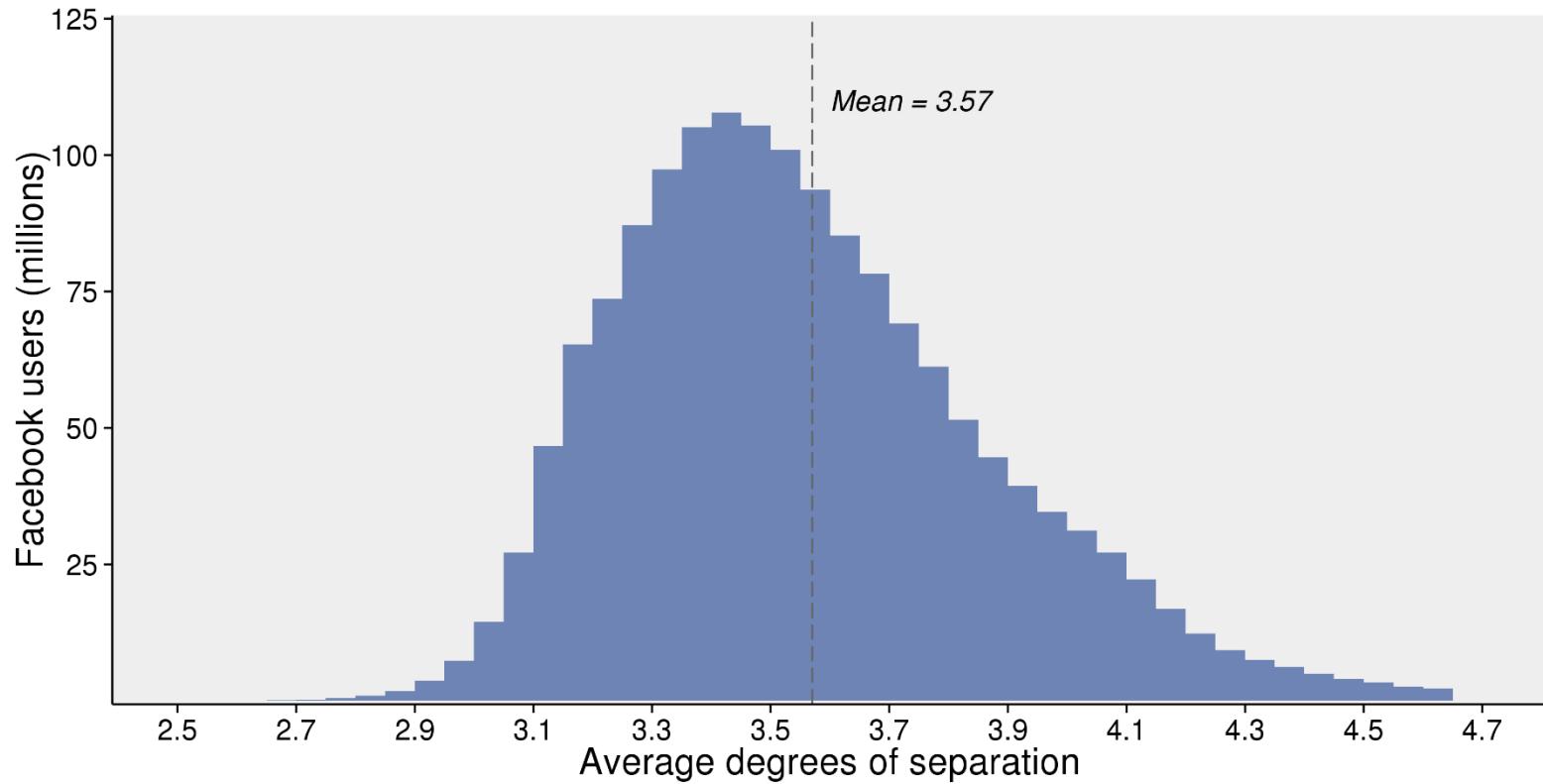
2

Experimentos posteriores

- ◆ P.e. Dodds, Muhamad & Watts 2003 repitieron el experimento a gran escala
- ◆ 18 personas objetivo en 13 países diferentes puntos del globo
- ◆ 24.163 cadenas de email, 384 llegan al destino
- ◆ Distancia promedio $\sim 4\text{-}6$



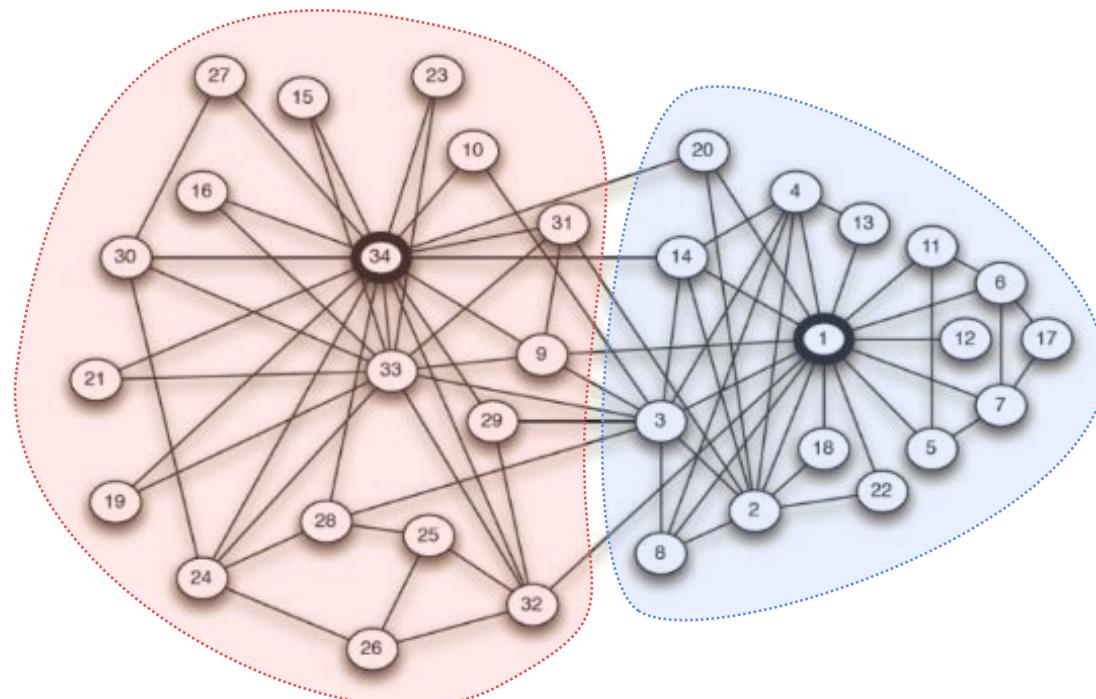
Small world en Facebook

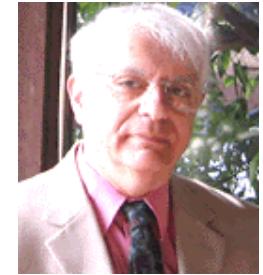


<https://research.facebook.com/blog/three-and-a-half-degrees-of-separation> (Feb 2016)

Zachary's karate club

- ◆ W. W. Zachary, 1970
- ◆ 34 miembros de un club de karate
- ◆ 78 relaciones de amistad fuera de las clases
- ◆ Ruptura de pequeños grupos: maximum flow – minimum cut
Ford–Fulkerson algorithm from 1 to 34

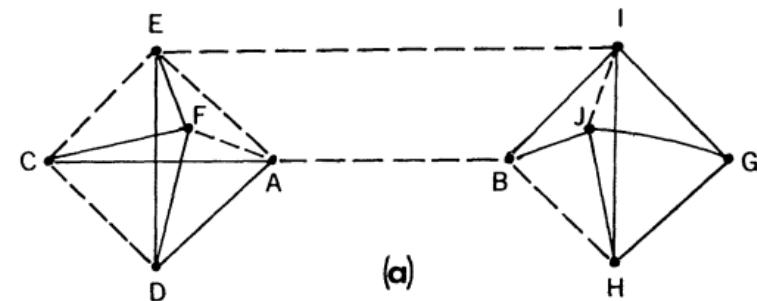




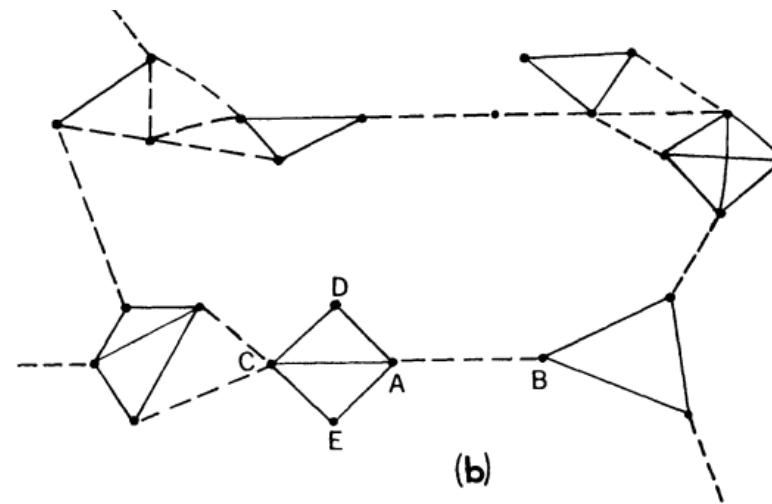
The strength of weak ties

◆ M. S. Granovetter, 1973

Mark S. Granovetter
(1943-)



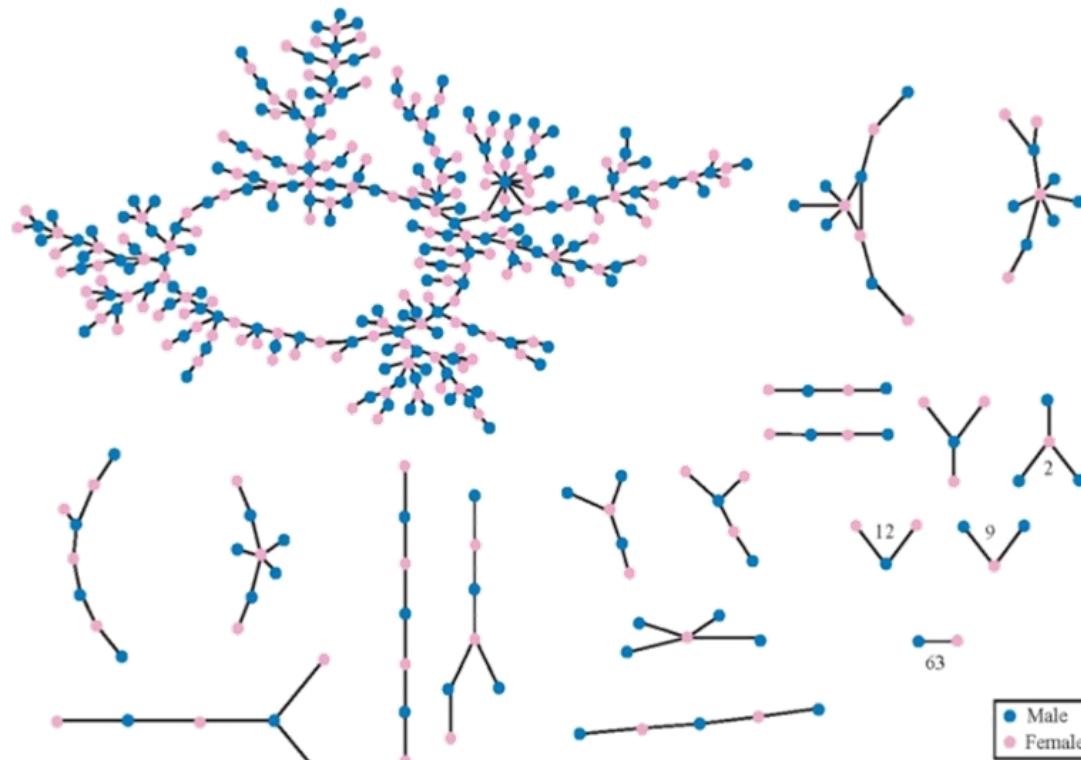
(a)



(b)

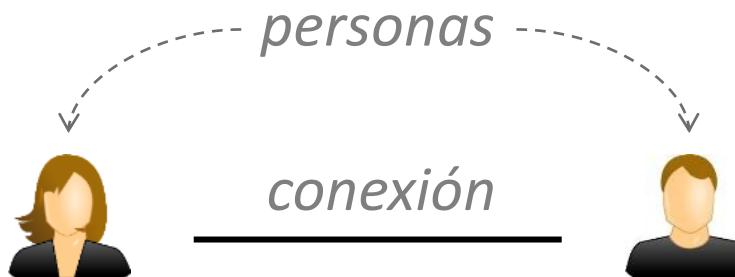
Estudio “Jefferson High”

- ◆ P. Bearman, J. Moody and K. Stovel, 1993
- ◆ 573 estudiantes ◆ Grado promedio 1.66
- ◆ ASP ~ 16, heterofilia, ciclos largos, bajo coeficiente de clustering
- ◆ Enfermedades de transmisión sexual



¿Qué es una red social?

La pieza básica:

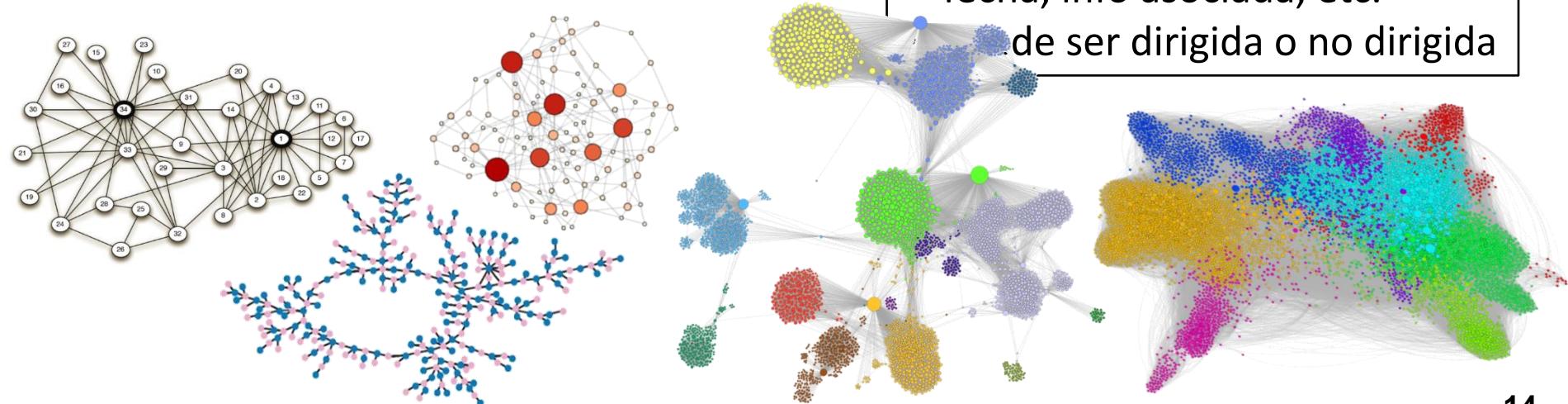


En principio personas, pero se puede generalizar a organizaciones, países, etc.

- Puede reflejar diferentes hechos: amistad, colaboración, comunicación, compartir, etc.
- Puede tener estructura: tipo, fecha, info asociada, etc.

de ser dirigida o no dirigida

Formación de estructuras complejas



Redes sociales

- ◆ Son un caso particular de los fenómenos de conectividad (*netwok science*)
 - Mucho en común con otras redes: de comunicación (transporte, Internet, etc.), relaciones y cadenas comerciales (países, empresas, organizaciones...), redes biológicas (cadena trófica, neuronas, interacción celular, etc.), interacciones químicas, grafos léxicos, la Web, etc.
- ◆ Las redes sociales pueden ser de muy diverso tipo
 - Amistad, colaboración, afiliación, parentesco, cita bibliográfica, interacción (email, teléfono, retweet, etc.)
 - Las redes online son un subconjunto visible de las redes completas
- ◆ Un campo multidisciplinar
 - Sociólogos, antropólogos, economistas, matemáticos, físicos, informáticos...

Breve perspectiva histórica

- ◆ Se empieza a hablar de redes sociales a finales del XIX
 - ◆ Primeros estudios en los 30 (sociogramas)
 - ◆ Formalización matemática en los 50
 - ◆ Amplio desarrollo de teorías y métodos en los 80
 - ◆ Trabajo de los físicos en los 90 (importantes avances teóricos)
 - ◆ Redes online en los 2000!
- 
- Ciencias sociales

“In the absence of
actual network data,
all this is speculation”

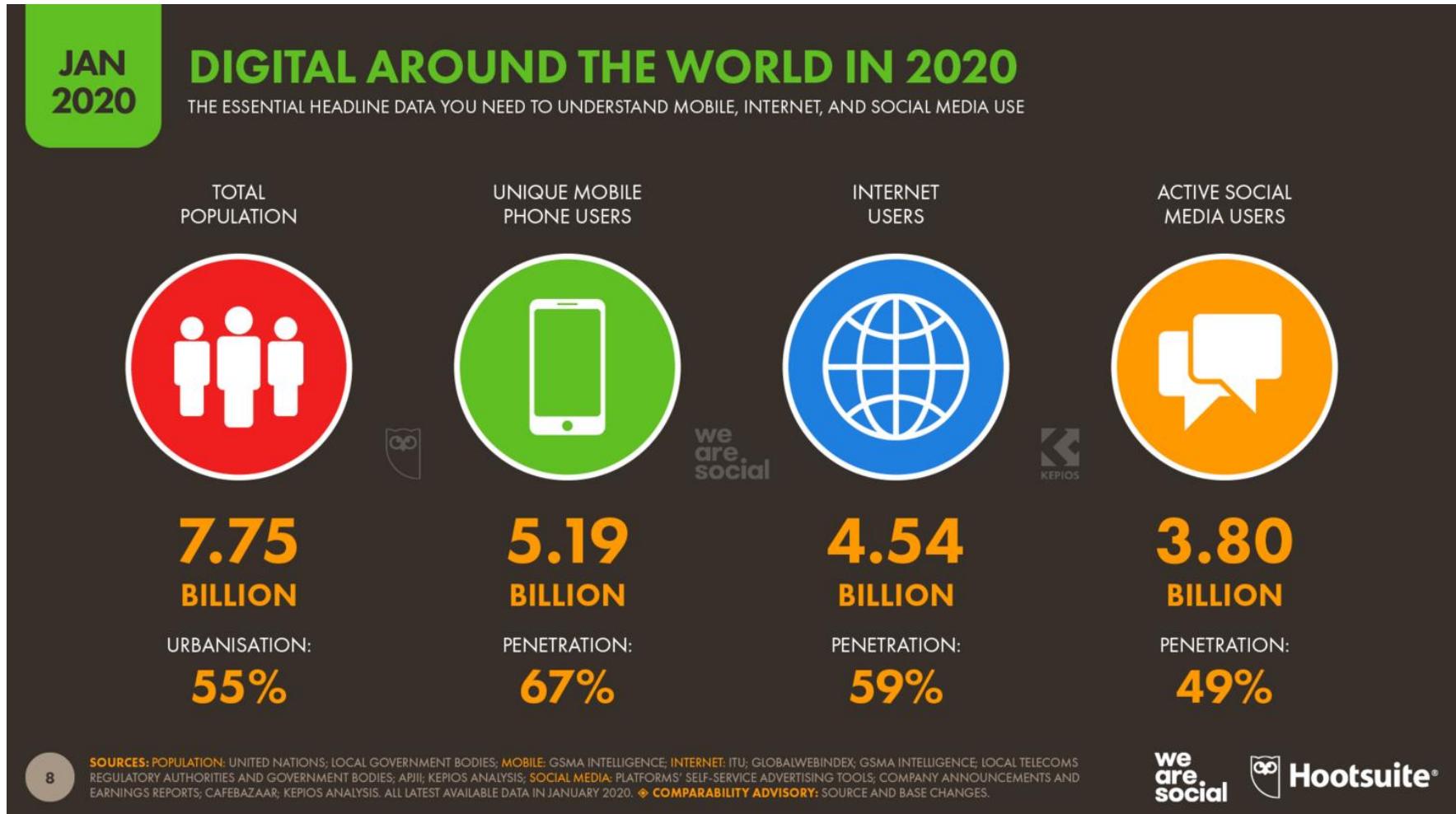


M. S. Granovetter. The strength of weak ties.
American Journal of Sociology 78(6), 1973

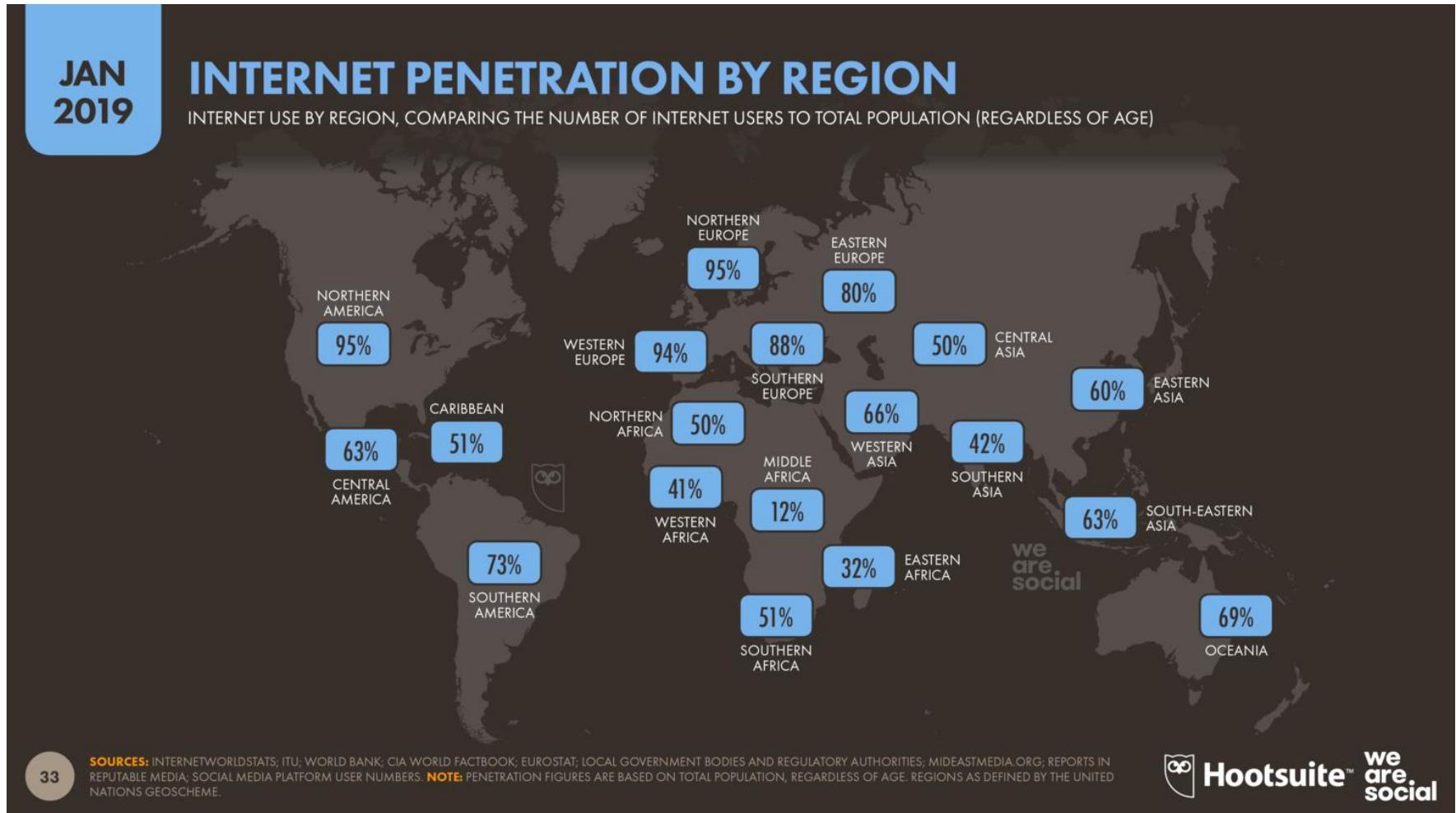
Redes sociales online

- ◆ Las redes online en los 2000 dan lugar a un boom de actividad en este campo
 - SixDegrees en 1997, Friendster en 2002, MySpace y LinkedIn en 2003, Orkut y Facebook en 2004, Twitter en 2006, Google+ en 2011...
- ◆ Tendencia a la integración de la red en plataformas de servicio

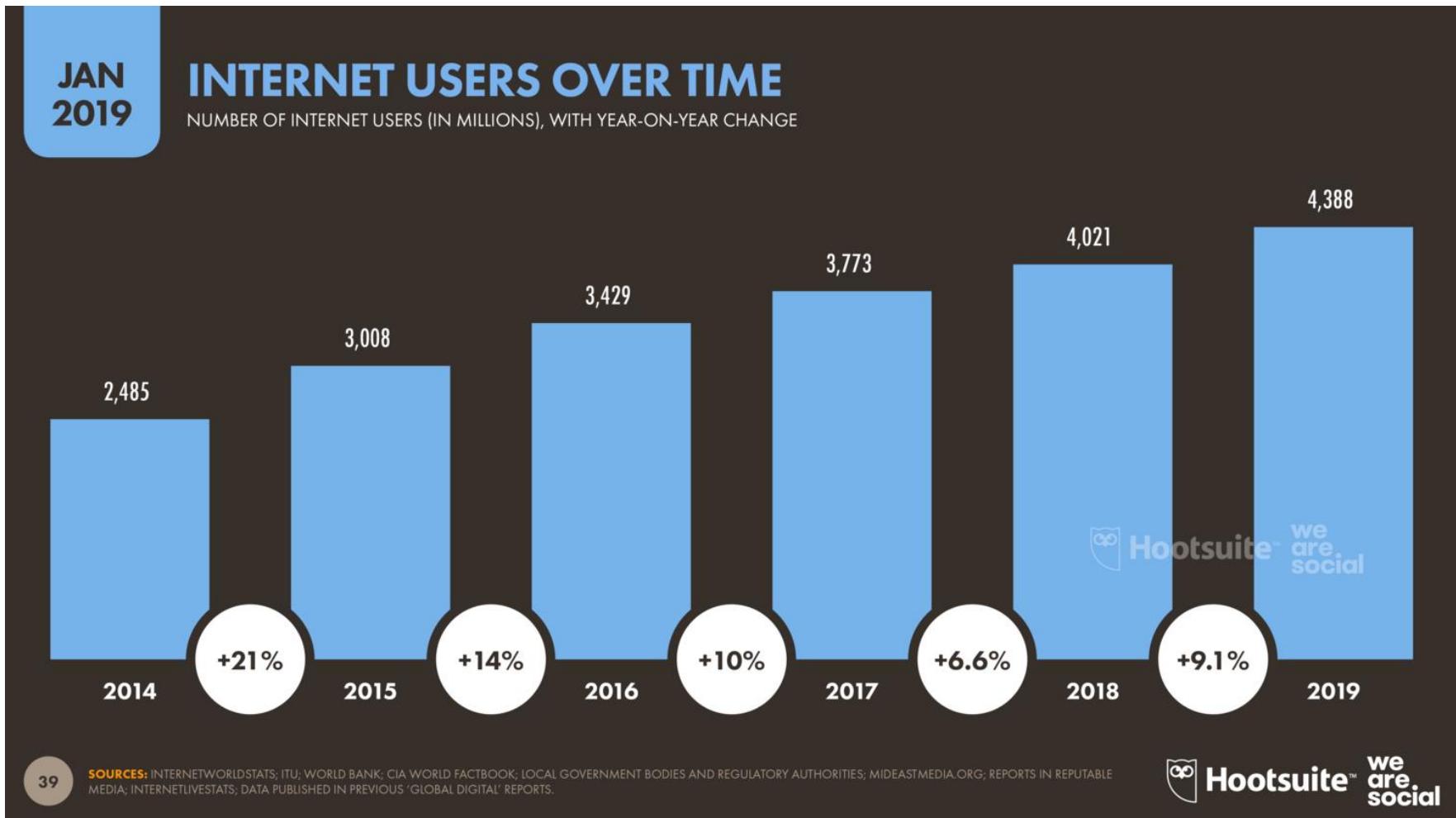
Cifras globales



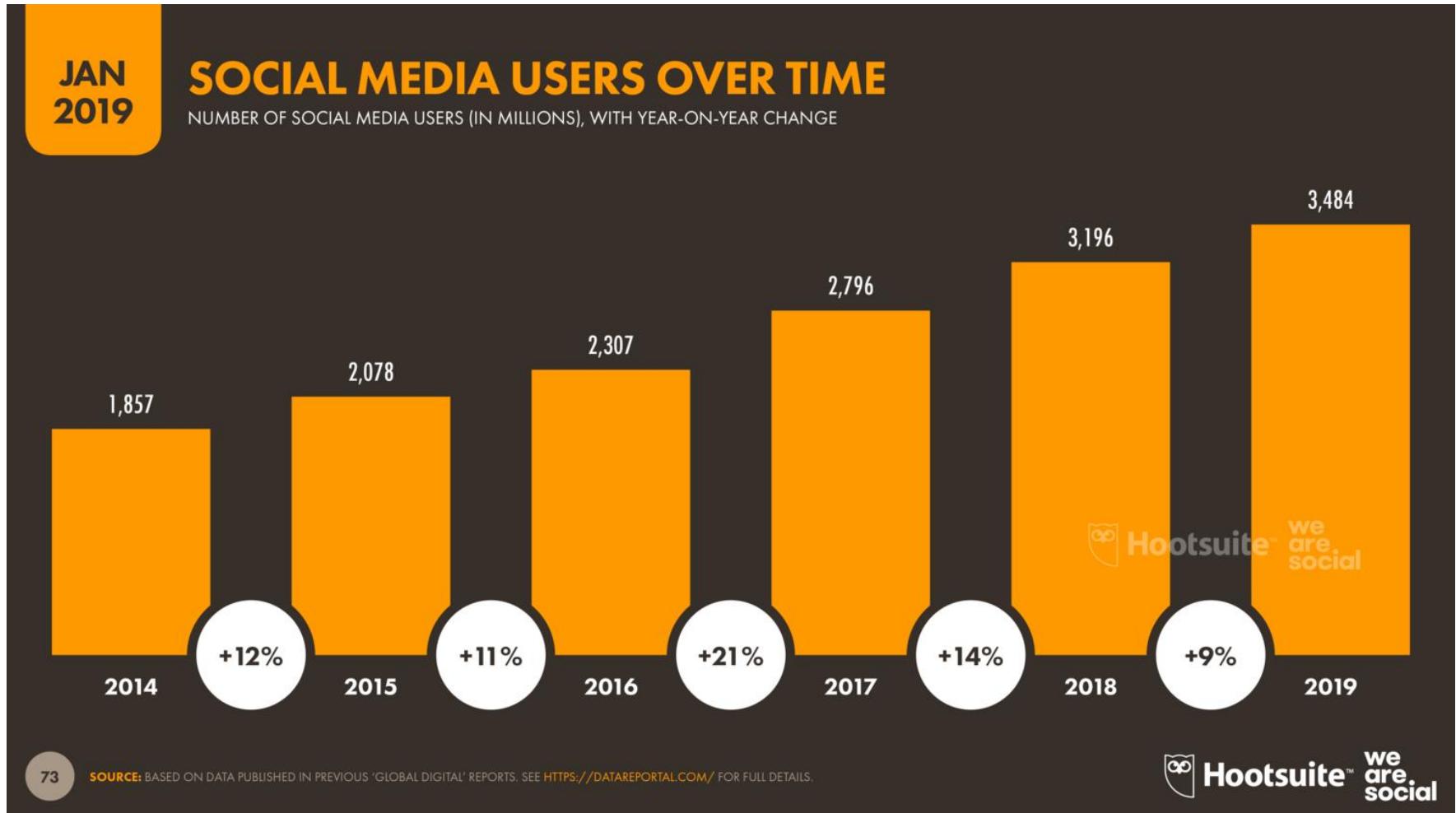
Cifras globales



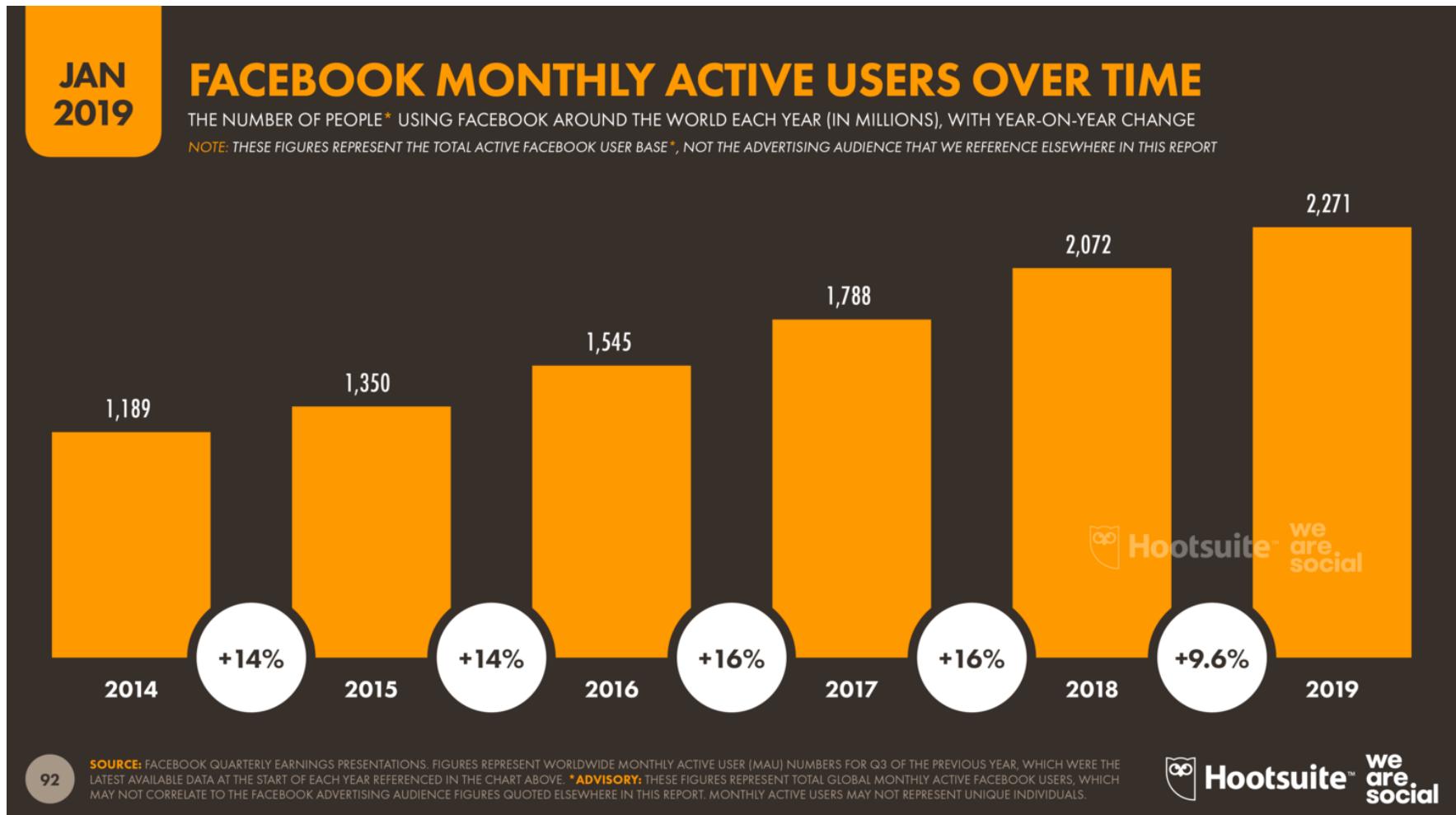
Cifras globales



Cifras globales



Cifras globales

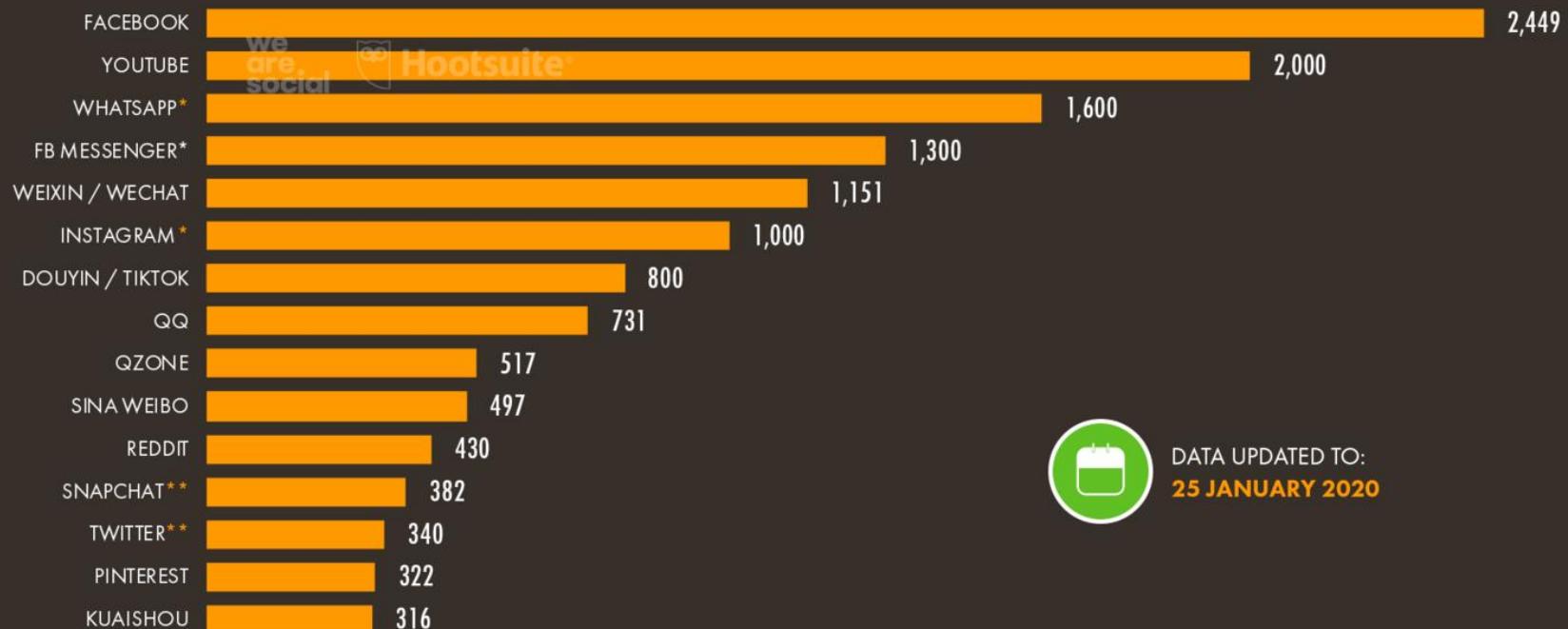


Cifras globales

JAN
2020

THE WORLD'S MOST-USED SOCIAL PLATFORMS

BASED ON MONTHLY ACTIVE USERS, ACTIVE USER ACCOUNTS, ADVERTISING AUDIENCES, OR UNIQUE MONTHLY VISITORS (IN MILLIONS)



DATA UPDATED TO:
25 JANUARY 2020

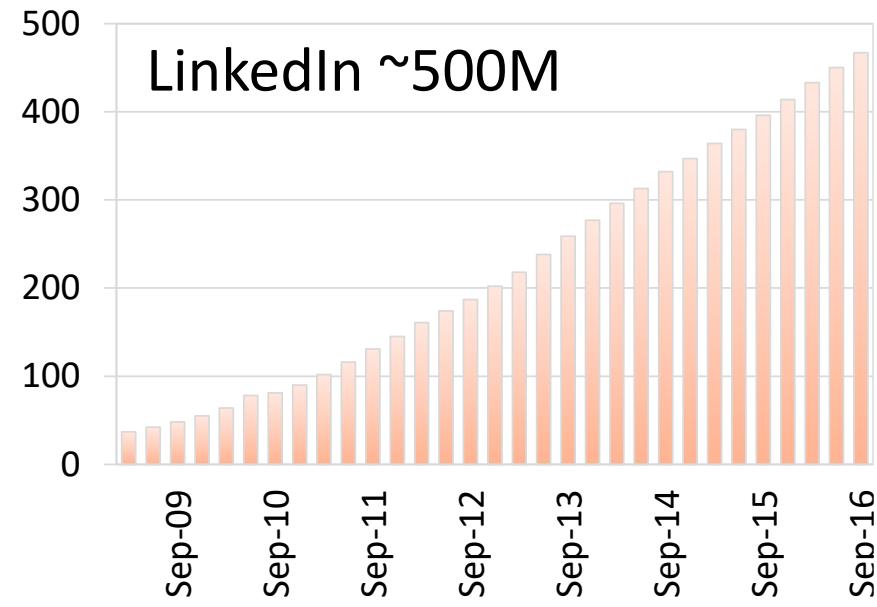
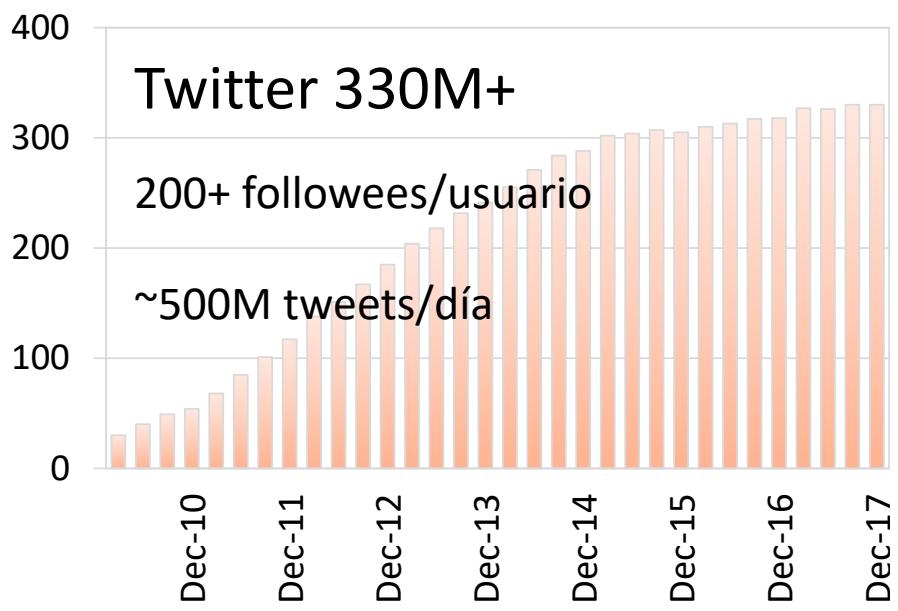
95

SOURCES: KEPiOS ANALYSIS; COMPANY STATEMENTS AND EARNINGS ANNOUNCEMENTS; PLATFORMS' SELF-SERVICE ADVERTISING TOOLS (ALL LATEST AVAILABLE DATA). **NOTES:** PLATFORMS IDENTIFIED BY (*) HAVE NOT PUBLISHED UPDATED USER NUMBERS IN THE PAST 12 MONTHS. PLATFORMS IDENTIFIED BY (**) DO NOT PUBLISH MAU DATA. FIGURES FOR TWITTER AND SNAPCHAT USE EACH PLATFORM'S LATEST ADVERTISING AUDIENCE REACH, AS REPORTED IN EACH PLATFORM'S SELF-SERVICE ADVERTISING TOOLS (JANUARY 2020).

**we
are
social**  **Hootsuite®**

Crecimiento de las redes sociales online

- ◆ Cientos de sitios (Wikipedia enumera 14 con 100M+ usuarios)
- ◆ Miles de millones de usuarios activos

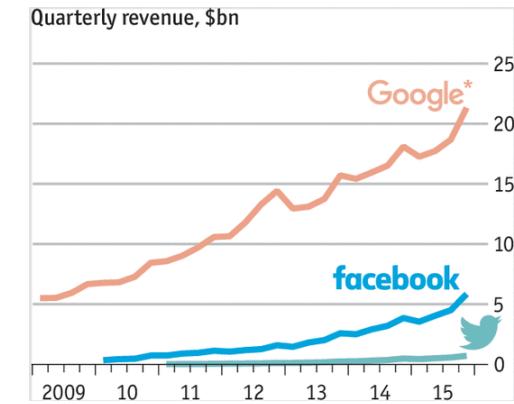


Volumen de negocio en las redes sociales

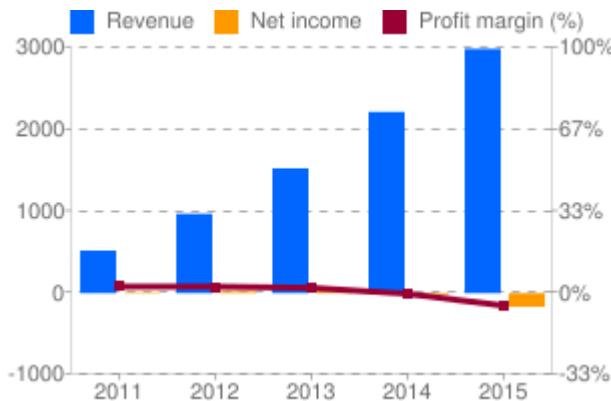
Facebook (17K+ empleados)



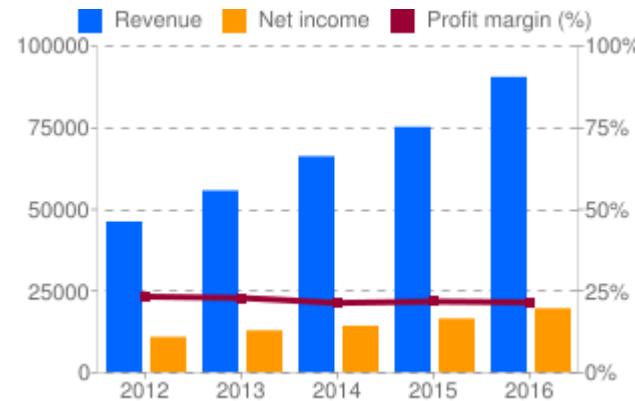
Twitter (~3.9 empleados)



LinkedIn (9.7K+ empleados)

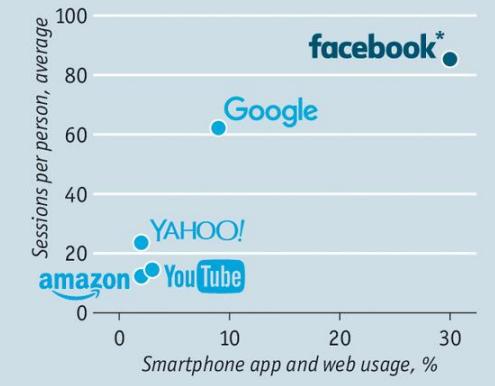


Google (70K+ empleados)



Tip-top tap

Smartphone usage in America, aged over 18
December 2015

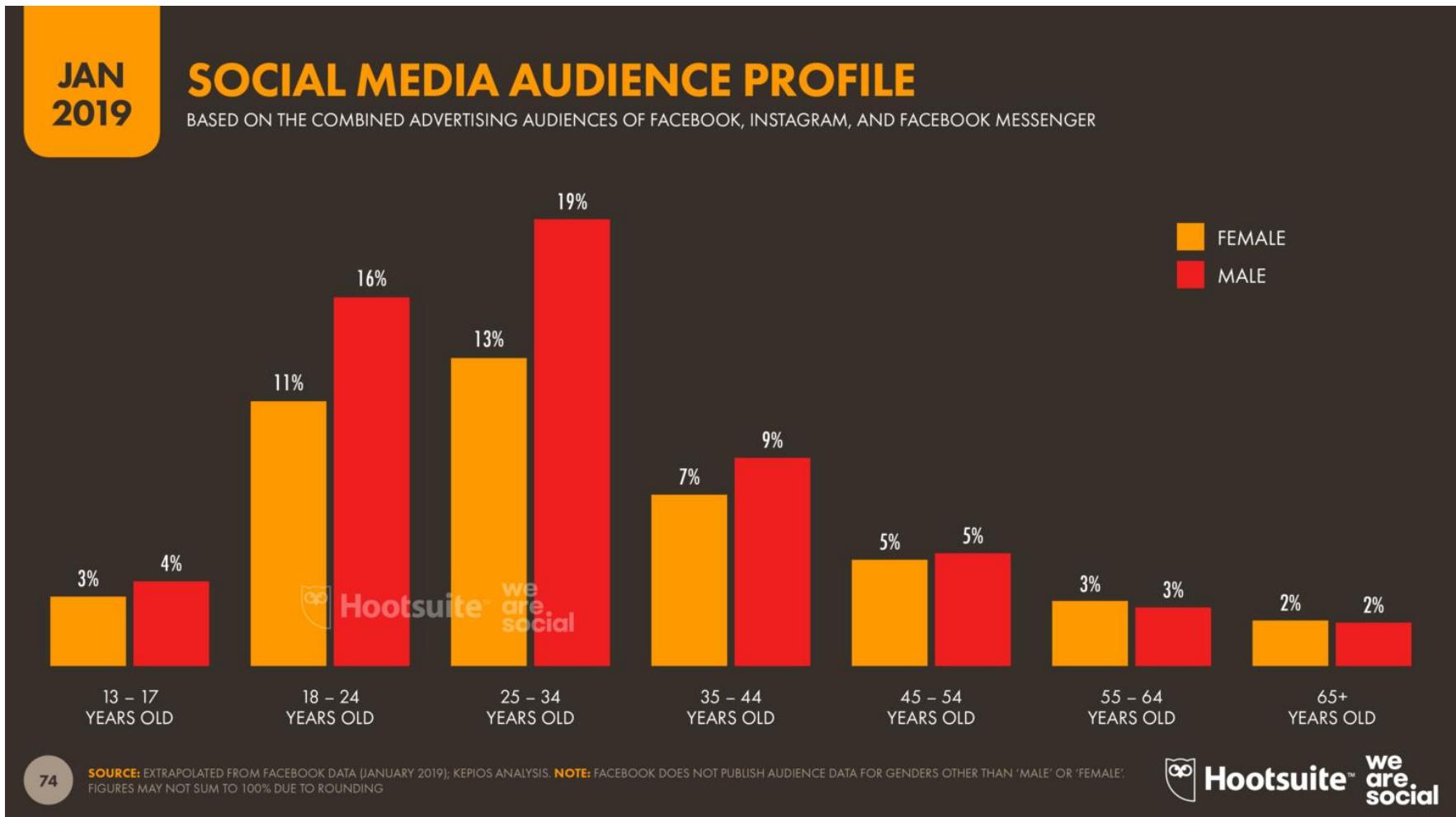


Source: Nielsen

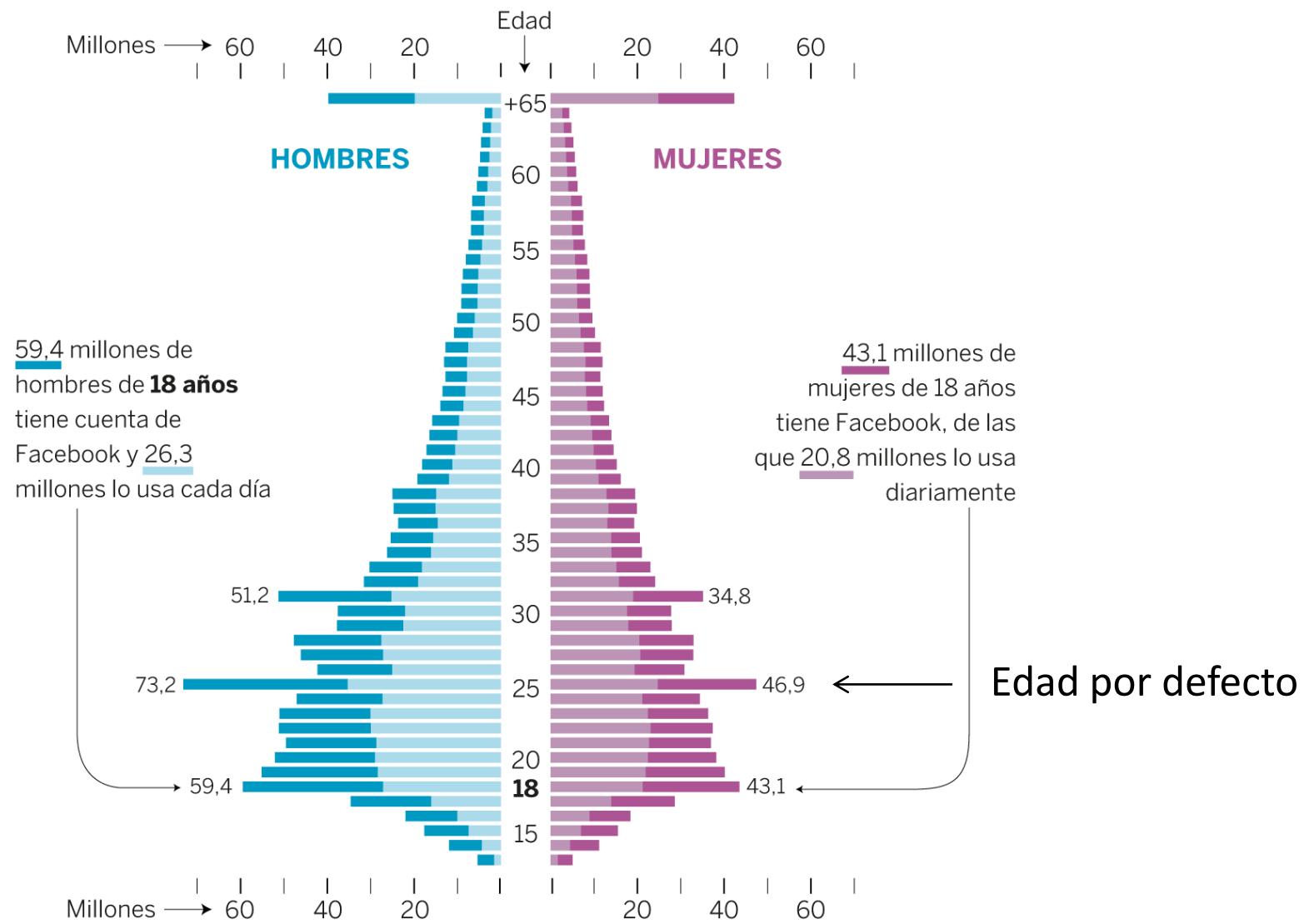
*Including Instagram and WhatsApp

Economist.com

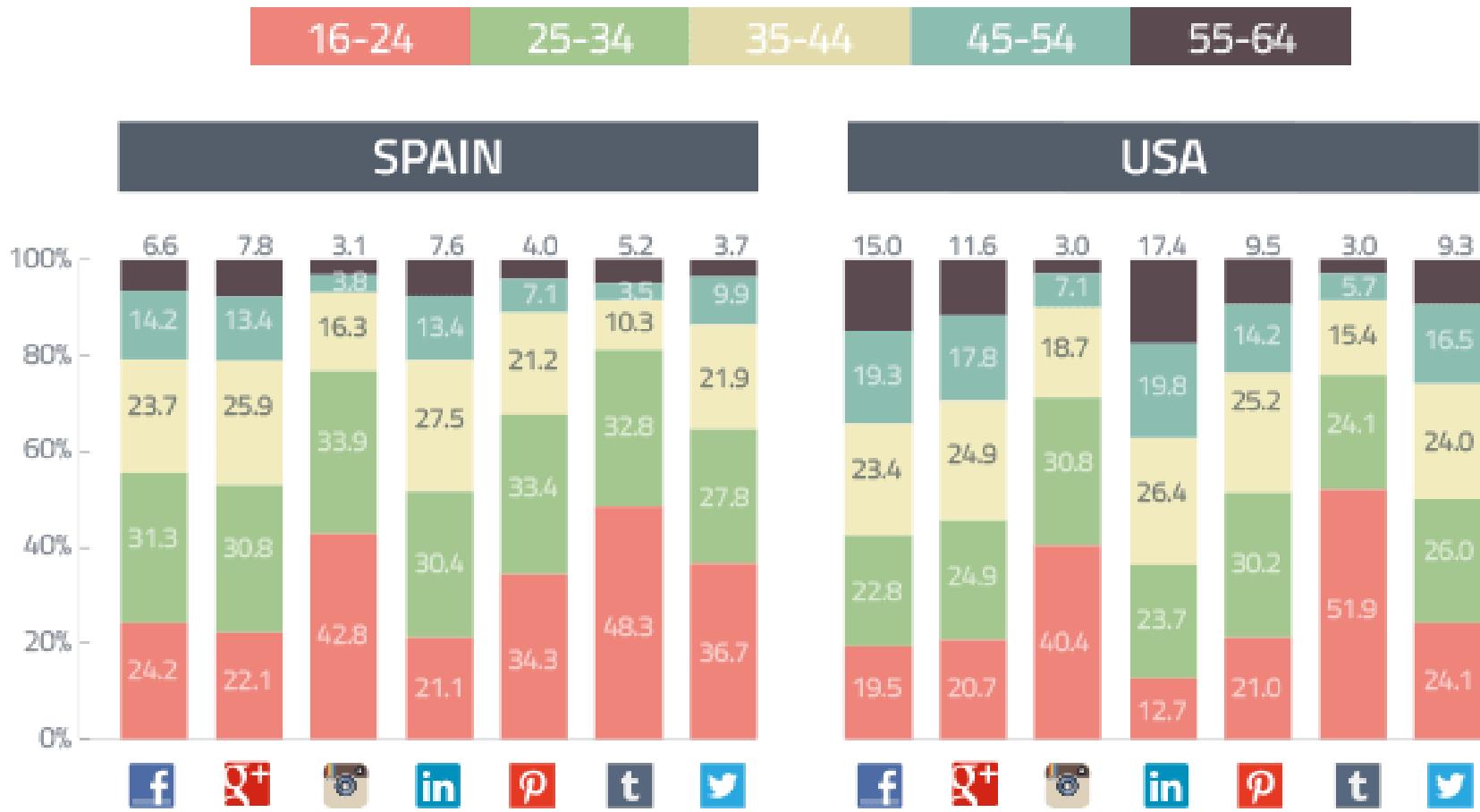
Cifras globales



Pirámide poblacional de Facebook



Distribución poblacional en medios sociales



Redes sociales (cont)

- ◆ Un campo muy abierto en potencial y soluciones
- ◆ Ejemplos de problemas de interés
 - A qué personas dirigir una campaña de publicidad; cómo ganar influencia en la red; descubrir, predecir, recomendar relaciones; cómo frenar una epidemia o maximizar la difusión de un mensaje; entender cómo se relacionan las personas (cómo eligen relaciones, cómo deciden asociarse, cómo interactúan, etc.); manipular unas elecciones 😬

Y con vistas a abordar tales problemas...

- ◆ Qué características nos interesa observar y analizar en una red social
 - P.e. aquéllas que determinan el comportamiento o respuesta global de la red ante estímulos: alcance y velocidad de la propagación de estados (información, enfermedades, opinión, decisiones), impacto de la eliminación de nodos y/o arcos, etc.
 - P.e. aquéllas que determinan el papel de ciertas personas o grupos: en procesos de difusión, influencia, toma de decisiones (comprar, votar, entrar en un grupo, asistir a un evento), etc.
- ◆ Explicar cómo y por qué se llegan a formar las características de una red, predecir su futuro desarrollo y evolución
- ◆ Explicar y predecir cómo se va a desarrollar un proceso (p.e. difusión) sobre la red

Redes sociales y minería/búsqueda de información

- ◆ Las redes sociales están cambiando la forma en que accedemos a la información
 - Flujo, compartición, propagación, búsqueda, acceso
 - Micro escala (nuestro entorno cercano) y macro escala (propagación viral, etc.)
- ◆ Se está empezando a descubrir la utilidad de las redes en algoritmos de recomendación
 - Nos pueden ser útiles actividades y hallazgos de nuestros amigos
- ◆ “Crowd power”
 - Se difumina la barrera entre productor y consumidor
 - Contenido pero también datos, estructura, respuestas, etc.
- ◆ Las personas son también objeto de tareas de búsqueda y análisis
 - Buscar/recomendar personas, expertos, candidatos a un trabajo, amigos, pareja, etc.
 - Analizar sus propiedades y relaciones: quién es influyente, qué comunidades se observan, qué factores determinan las interacciones, etc.

Análisis de redes sociales

- ◆ Medir y describir
 - Métricas
- ◆ Explicar
 - Modelos
- ◆ Predecir
 - Anticipar fenómenos
 - Influir en ellos
- ◆ Aspectos estáticos
 - Topología (local y global)
 - Propiedades de personas y enlaces individuales
- ◆ Aspectos dinámicos
 - Aparición y desaparición de enlaces
 - Formación y variación de propiedades
 - Interacción: flujo de información y estados

¿Qué se estudia sobre las redes sociales?

- ◆ Topología micro, meso, macro
 - Componentes conexas, densidad, cohesión, distancias...
 - Redes de mundo pequeño, ley de potencias
 - ◆ Propiedades estructurales de los nodos (y enlaces)
 - Cómo están posicionadas las personas en su entorno y qué nodos son “importantes”
 - Diferentes tipos de importancia: autoridad, centralidad, influencia, mediación...
 - Propiedades relacionadas con la cohesión del entorno
 - Enlaces fuertes / débiles
 - ◆ Modelos de estructura y formación de las redes
 - Tendencias elementales (micro) y topologías a las que dan lugar
 - Modelos de grafos “aleatorios”
 - ◆ Comunidades
 - ◆ Fenómenos de propagación
 - ◆ Optimización de algoritmos para grafos de muy alta escala
 - ◆ Visualización (un reto para escalas masivas)
 - ◆ ...
- 
- “Forma” de una red:
para escalas masivas
se mide en términos de
estadísticas y métricas
cuantitativas

2. Topologías de red: métricas y estadísticas

Nociones generales de grafos

- ◆ Grafo $G = (V, A)$, nodos $u \in V$, arcos $(u, v) \in A$, $|V| = n$, $|A| = m$
- ◆ $g(u) \equiv$ grado de un nodo $u \in V$
 - En grafos dirigidos indegree, outdegree
- ◆ Tipos de grafos
 - Dirigidos / no dirigidos → por defecto vamos a suponer grafos no dirigidos, salvo cuando digamos expresamente lo contrario
 - Ponderados / no ponderados
 - Se podrían considerar multigrafos
- ◆ Componentes (fuertemente) conexas
- ◆ Caminos de distancia mínima (a.k.a. geodésicas)
 - Y de coste mínimo en grafos con arcos ponderados
- ◆ Red ego de un nodo: subgrafo a distancia ≤ 1 del nodo

A qué aplican las métricas y el análisis

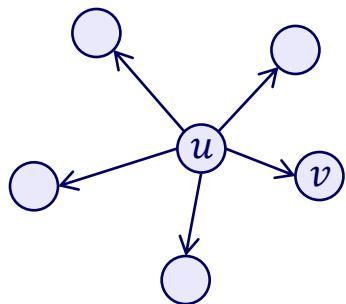
- ◆ Nodos individuales
- ◆ Arcos individuales
- ◆ La red en su totalidad

Propiedades de los nodos

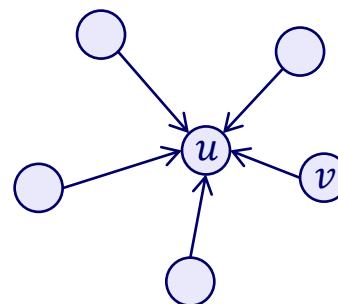
- ◆ Métricas que miden propiedades topológicas de nodos individuales, relacionadas con su importancia y el papel que pueden jugar en la red
- ◆ No hay un conjunto canónico de métricas
 - Se han definido cientos, veremos las más fundamentales y conocidas
 - Pero frecuentemente se “inventan” nuevas métricas para problemas particulares
- ◆ Grado
 - Distinción entre grado / indegree / outdegree en redes dirigidas
- ◆ Centralidad: betweenness, closeness, PageRank, autovector
- ◆ Cohesión local: coeficiente de clustering
- ◆ Existen ligeras variantes en la definición de algunas de estas métricas
 - Diferentes formas de normalizar, excluir o no el nodo en las sumas, etc.
 - Las diferencias son generalmente intrascendentes en tanto que preservan las comparaciones fijada una red (y en algunos casos para toda red)

Ejemplos

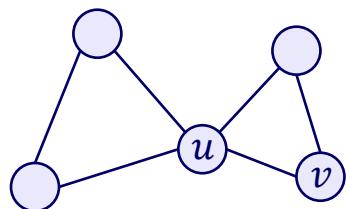
Outdegree



Indegree



Betweenness



Closeness



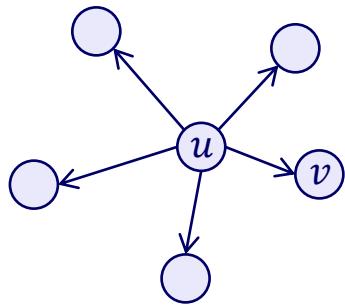
En todos estos ejemplos la propiedad es mayor en u que en v

Grado

- ◆ Nº de enlaces (en redes dirigidas, salientes / entrantes) en los que participa un nodo
- ◆ Es la métrica más simple, pero no menos significativa
 - Ya nos dice algo sobre el papel y/o importancia de los nodos
- ◆ Se considera a menudo una métrica de centralidad (es común que tenga relación p.e. con la influencia del nodo)
- ◆ Es relevante estudiar asimismo la distribución del grado como uno de los elementos característicos de una red

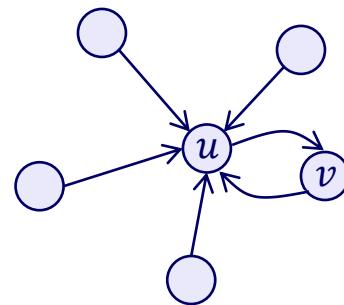
Grado

Outdegree



	g_{in}	g_{out}
u	0	5
v	1	0

Indegree

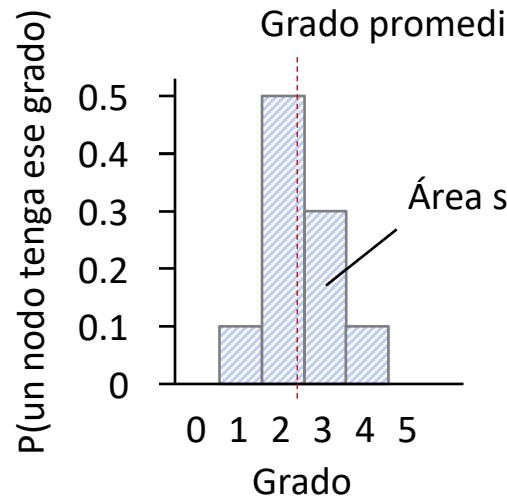
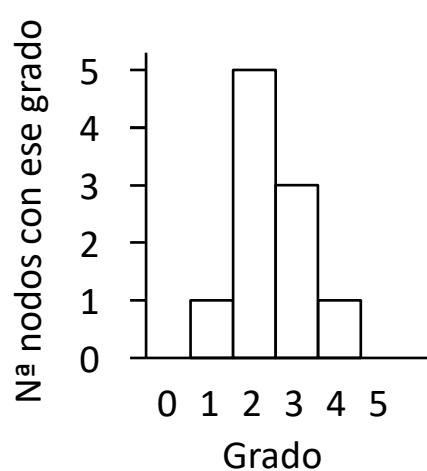
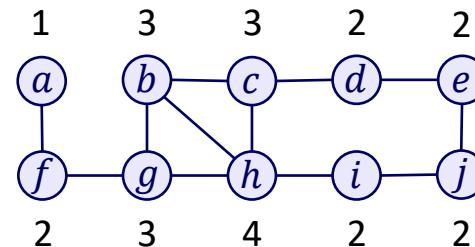
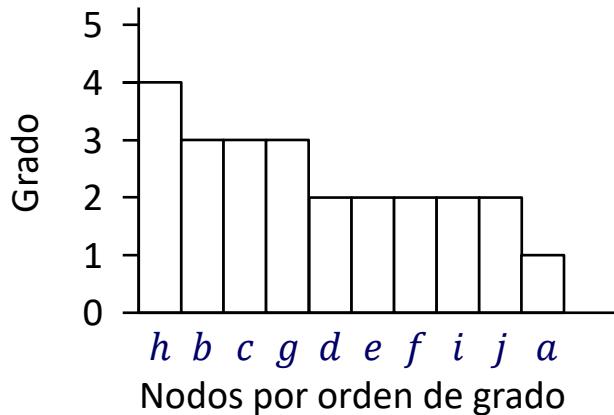


	g_{in}	g_{out}	g
u	5	1	5
v	1	1	1

Distribución del grado

- ◆ Forma parte esencial de la visión global de la estructura de una red
- ◆ La distribución puede visualizarse como una serie numérica: los grados ordenados de mayor a menor
- ◆ O bien, más comúnmente, se observa la frecuencia de los grados (cuántos nodos de grado 1, cuántos de grado 2, etc.)
 - Equivalente a la función de masa de la probabilidad de que un nodo al azar tenga un cierto grado (la diferencia está en dividir o no por el nº de nodos)
- ◆ Las distribuciones en redes naturales suelen estar típicamente muy sesgadas, como veremos
- ◆ También se estudia la distribución del grado en redes modelo (aleatorios), típicamente se consigue derivar una fórmula exacta

Ejemplo



Ejemplo

Grafo Facebook (J. Leskovec)

Red ego 10 usuarios

$$|V| = 4,039$$

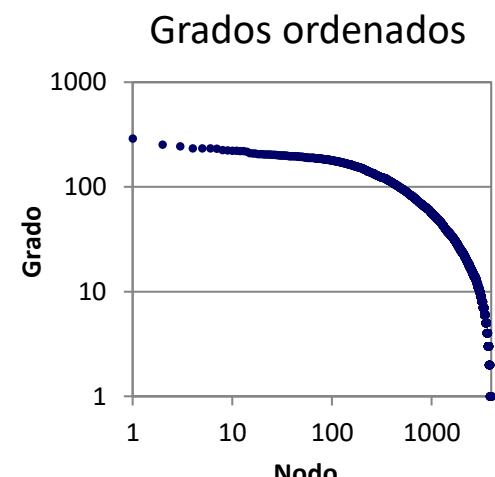
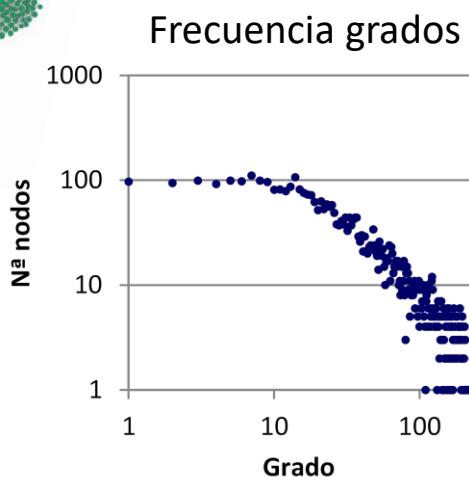
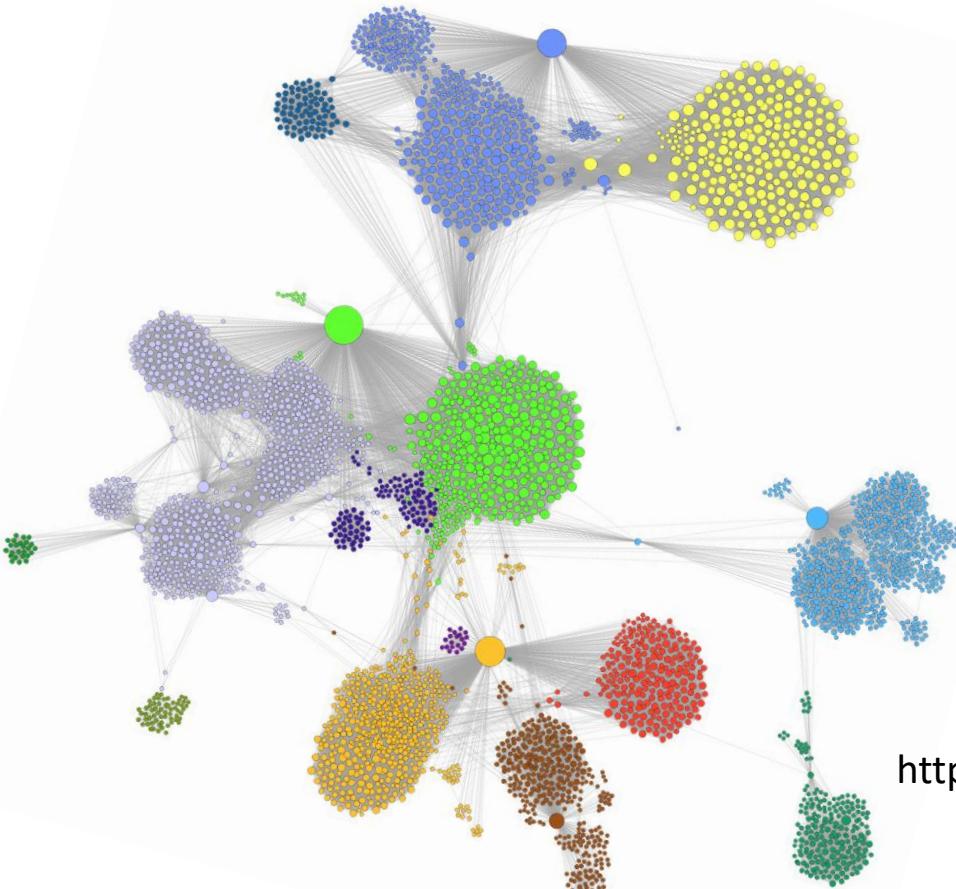
$$\text{avg}_u g(u) = 43.7$$

$$C_{\text{avg}} = 0.617$$

$$ASP = 3.7$$

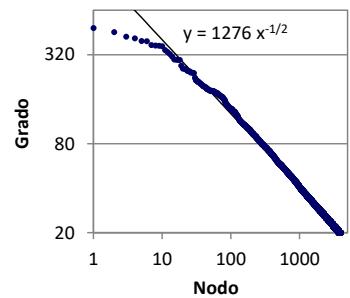
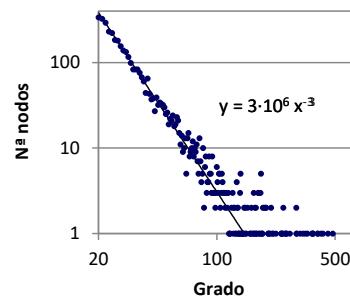
$$\text{Diámetro} = 8$$

<http://snap.stanford.edu/data/egonets-Facebook.html>



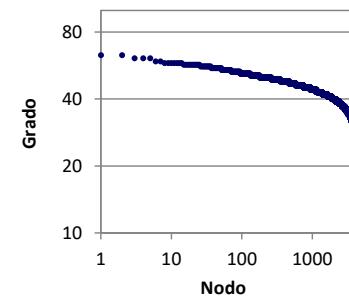
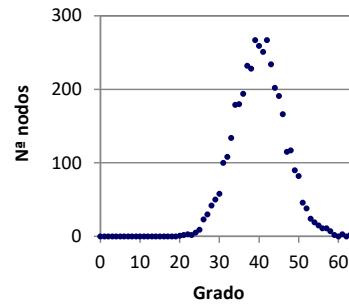
Más ejemplos

Grafo Barabási-Albert
 $|V| = 4,000$
 $\text{avg}_ug(u) = 40$
 $ASP = 2.56 (3.9)$
 $\text{Diámetro} = 4$
 $C_{\text{avg}} = 0.036 (0.002)$

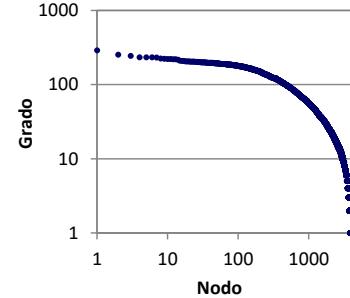
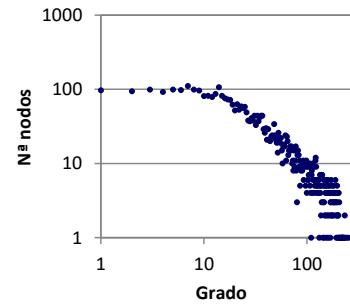


Redes generadas por modelo

Grafo Erdös-Rényi
 $|V| = 4,000$
 $\text{avg}_ug(u) = 40$
 $ASP = 2.65 (2.76)$
 $\text{Diámetro} = 4$
 $C_{\text{avg}} = 0.01 (0.01)$



Grafo Facebook (J. Leskovec)
 Red ego 10 usuarios
 $|V| = 4,039$
 $\text{avg}_ug(u) = 43.7$
 $ASP = 3.7$
 $\text{Diámetro} = 8$
 $C_{\text{avg}} = 0.617$



Red datos reales

Paradojas de la amistad

- ◆ ¿Mis amigos tienen más amigos que yo?
- ◆ El promedio del nº de amigos de los amigos es mayor que el nº promedio de amigos por persona

$$\text{avg}_u g(u) \leq \text{avg}_{u,v:u \rightarrow v} g(v)$$

- Es un hecho estadístico fácil de comprobar
- Intuición: el grado de las personas con muchos amigos participa más veces en la suma que forma el promedio
- O bien: es estadísticamente más probable ser amigo de alguien con muchos amigos que de alguien con pocos

Paradojas de la amistad (cont)

- ◆ También se cumple siempre $\text{avg}_u g(u) \leq \text{avg}_u \text{avg}_{v:u \rightarrow v} g(v)$
 - Es una formulación ligeramente distinta, fácil de demostrar también
- ◆ No es teóricamente necesario sin embargo que la **mayoría** de personas tengan menos amigos que sus amigos
 - Es otra formulación distinta: que la **mediana** sea menor que la media $m < \mu$
 - Esto se cumple si la distribución del grado es monótona decreciente
 - Así suele ocurrir en las redes naturales
 - Sucedería lo contrario si los grados altos abundasen más que los bajos

Métricas de nodos basadas en distancias

- ◆ Closeness
- ◆ Excentricidad
- ◆ Betweenness

Closeness

- ◆ No necesariamente muchos contactos, ni punto de paso, pero en una posición cercana en promedio a todos los nodos

– Intuitivamente, estar “en medio” de la red en términos de distancia

- ◆ Se manejan variantes con ligeras diferencias, por ejemplo:

$$C(u) = \frac{n - 1}{\sum_{v \in V} \delta(u, v)} \text{ // inversa de la distancia mínima media}$$

- ◆ Refleja una posición de influencia por la rapidez para llegar a los demás nodos p.e. en el paso de información

- ◆ En redes naturales las distancias $\delta(u, v)$ suelen ser muy cortas, por lo que $C(u)$ varía poco entre nodos, y es una métrica inestable a pequeños cambios en la red

– P.e. con un solo enlace a un nodo muy central se dispara el valor

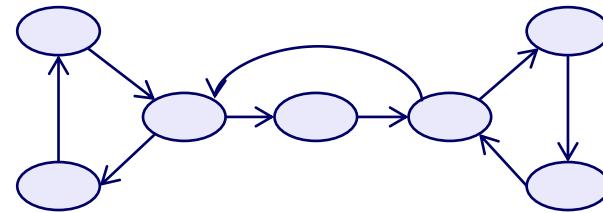
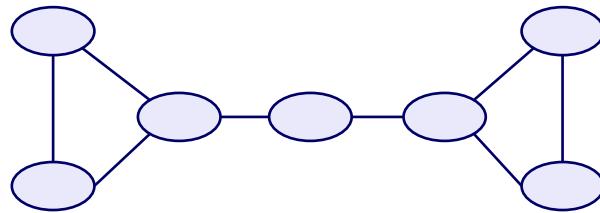
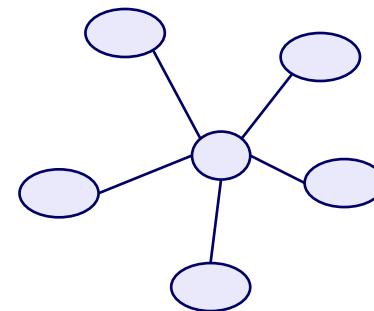
- ◆ Computación

– Para todos los nodos: calcular todos los CDMs! $O(n(n + m))$ // u $O(nm)$ Brandes
– Pero para un solo nodo, calcular un solo árbol CDM (con fuente en el nodo)

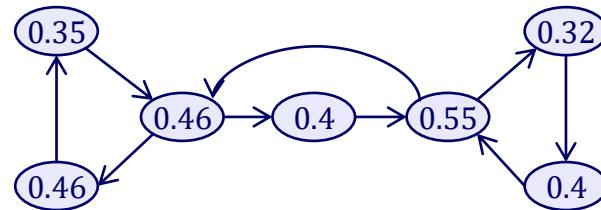
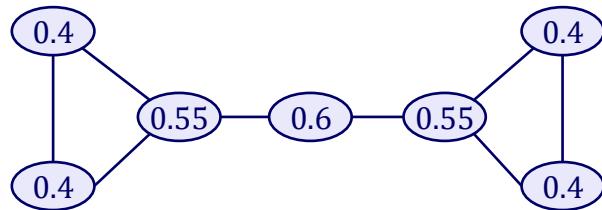
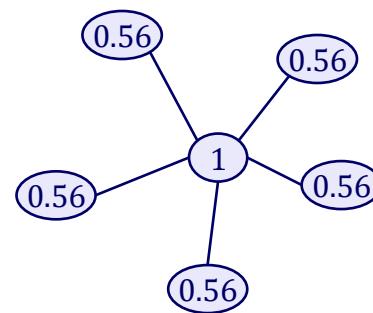
Closeness (cont)

- ◆ Cuando una red no es fuertemente conexa, todos los usuarios u tendrían $\delta(u, v) = \infty$ para algún v , y por tanto $C(u) = 0$
- ◆ Dos opciones para evitarlo
 - a) Calcular closeness en las componentes conexas como grafos separados
 - b) Closeness armónica: promedio de la inversa de las distancias en lugar de inversa de la distancia promedio: $C(u) = (n - 1) \sum_{v \neq u} 1/\delta(u, v)$

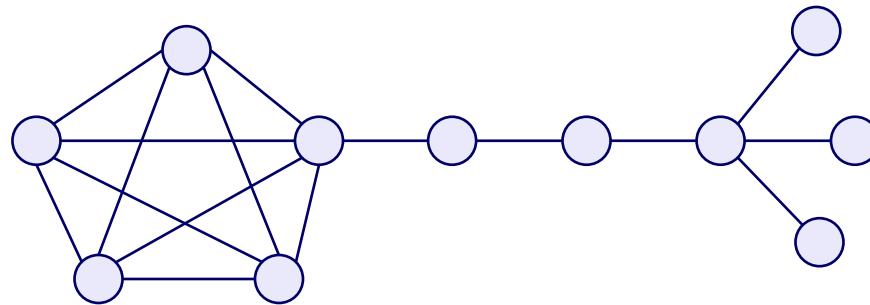
Ejemplos



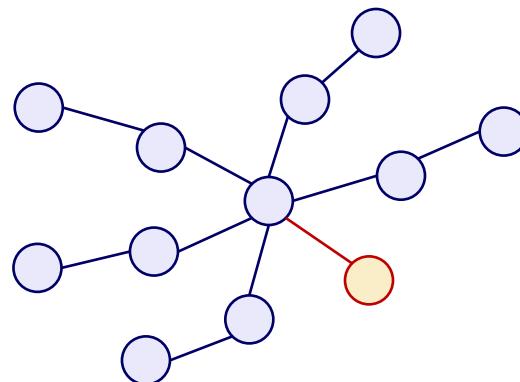
Ejemplos



Ejemplos (cont)



Grado no siempre implica closeness: este grafo contiene nodos con alto grado y bajo closeness, y viceversa. ¿Cuáles?



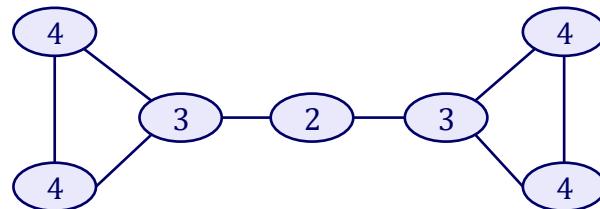
Closeness tampoco implica necesariamente betweenness

Excentricidad

- ◆ La distancia al nodo más lejano

$$e(u) = \max_{v \in V} \delta(u, v)$$

- ◆ Medida complementaria a closeness: en lugar de distancia mínima promedio, distancia mínima máxima



Betweenness

- ◆ No necesariamente muchos contactos, pero punto de paso entre muchos pares de nodos
- ◆ Ratio promedio de caminos de distancia mínima (CDM) de la red que pasan por el nodo

$$B(u) = \frac{2}{n(n-1)} \sum_{v,w \neq u} \frac{ns_{v,w}(u)}{ns_{v,w}}$$

En grafos dirigidos
omitimos esta condición

1 en grafos
dirigidos

2

$$\sum_{v,w \neq u} \frac{ns_{v,w}(u)}{ns_{v,w}}$$

$ns_{v,w}$ ≡ nº de CDM entre los nodos v y w

$ns_{v,w}(u)$ ≡ nº de CDM entre v y w que pasan por u

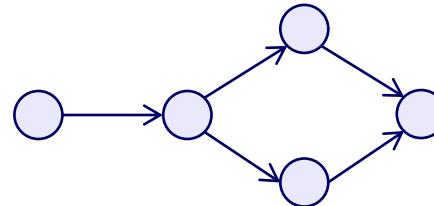
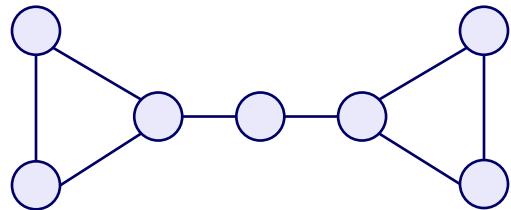
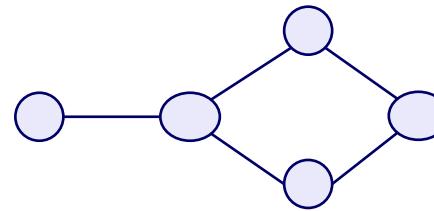
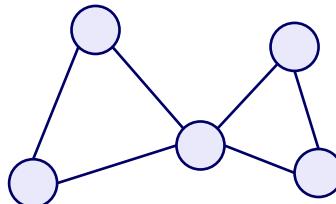
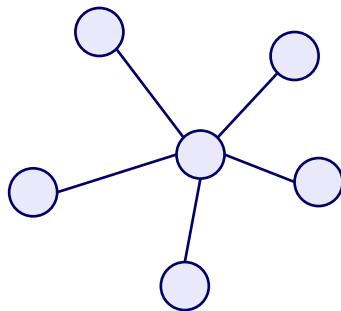
A menudo no se normaliza

Betweenness (cont)

- ◆ Si la red no es fuertemente conexa
 - Aplicamos la suma a los pares u, v tales que v es accesible desde u
 - Y normalizamos por el número de tales pares (o bien normalizamos dentro de cada componente fuertemente conexa)
- ◆ Cómputo: se necesita calcular todos los CDMs entre todos los pares de nodos (incluso para un solo nodo!)
 - BFS → árbol de CDMs desde cada nodo: $O(n(n + m))$
 - Una vez creado el bosque CDM, se calcula betweenness en $O(n(n + m))$
 - En grafos no dirigidos, $O(n m)$ con algoritmo de Brandes
 - En redes con pesos, Dijkstra / Johnson $O(n m + n^2 \log n)$, Floyd-Warshall $\Theta(n^3)$
- ◆ Los nodos con un valor alto en esta métrica tienen una posición de influencia por su papel en el paso de información
 - Su eliminación de la red tiende a crear disrupción del flujo de información

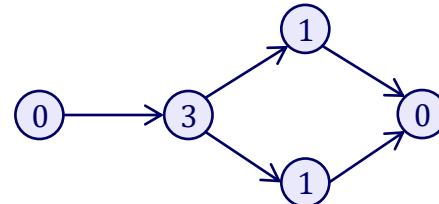
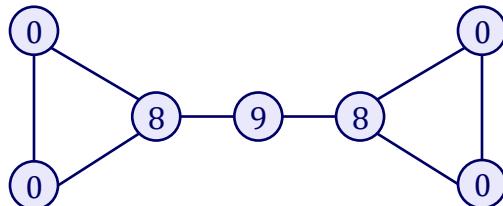
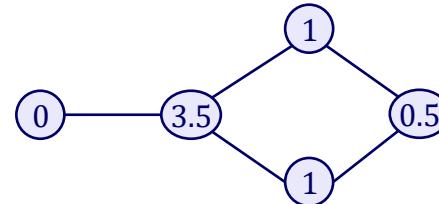
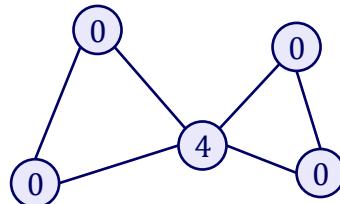
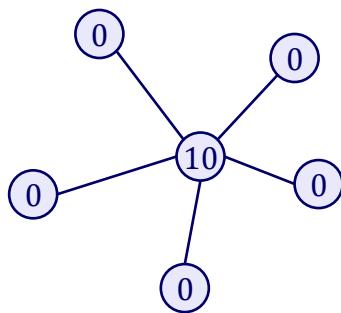
Ejemplos

(Valores sin normalizar)

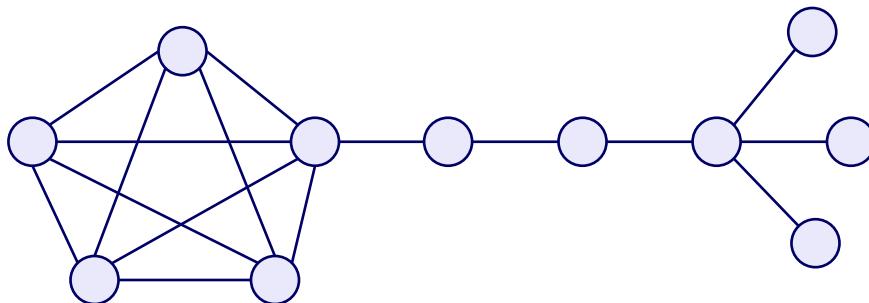


Ejemplos

(Valores sin normalizar)

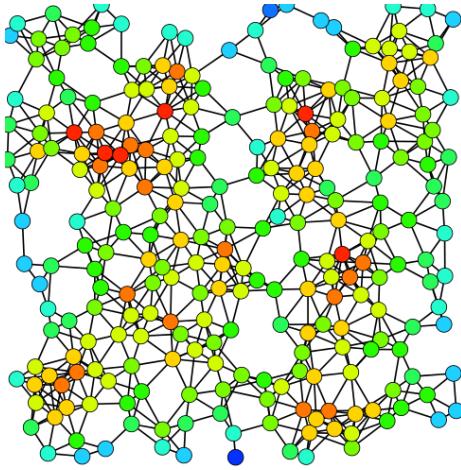


Ejemplos (cont)

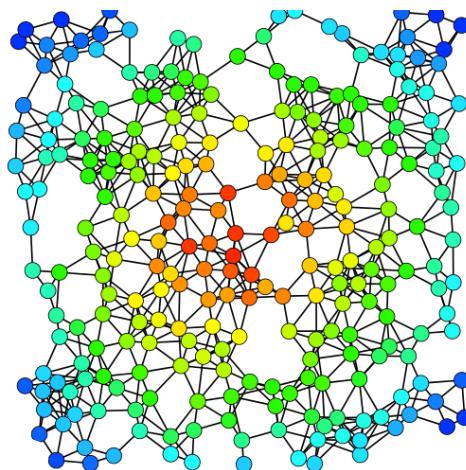


- ◆ Grado no siempre implica betweenness: este grafo contiene nodos con alto grado y bajo betweenness, y viceversa. ¿Cuáles?
- ◆ Betweenness tampoco implica closeness: ¿Qué nodos tienen alto betweenness y closeness moderado?

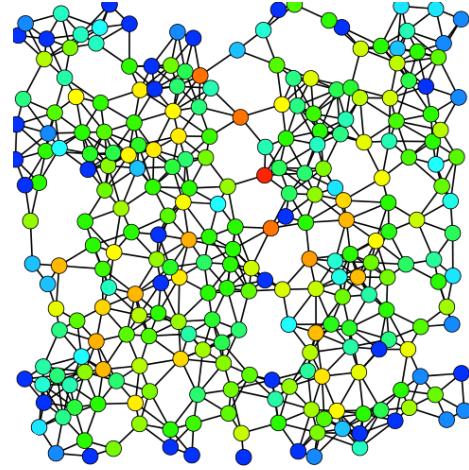
Comparación



Grado



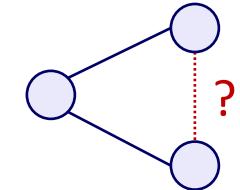
Closeness



Betweenness

Coeficiente de clustering local

- ♦ Refleja la cohesión del entorno de un nodo
- ♦ Se basa en la noción de cierre triádico
 - Transitividad: “los amigos de mis amigos son mis amigos”
 - En qué medida mis vecinos están conectados entre sí
 - Una forma de medir cómo de completo es el grafo entorno al nodo (“red ego”)
- ♦ $C(u) \equiv$ probabilidad de que dos vecinos de u tomados al azar sean vecinos



$$C(u) = p(v \rightarrow w | u \rightarrow v, u \rightarrow w) = \frac{\text{nº conexiones entre vecinos de } u}{\text{nº conexiones posibles entre vecinos de } u} \in [0,1]$$

$$g(u) < 2 \Rightarrow C(u) \triangleq 0$$

$$\text{nº conexiones posibles entre vecinos de } u = g(u)(g(u) - 1)/2$$

- ♦ El coef de clustering tiende a correlacionar inversamente con betweenness
 - Alto clustering \rightarrow redundancia en la comunicación
 - Bajo clustering \rightarrow posición ventajosa en la transmisión de información
 - Y es generalmente menos costoso de computar: $\Theta(\sum_{u \in V} g(u)^2)$
 - Aunque en redes power law muy sesgadas se puede disparar $\sum_{u \in V} g(u)^2$

Coeficiente de clustering global

- ◆ Refleja la cohesión global de entornos en la red
- ◆ Probabilidad de que dos nodos de la red con un amigo común tomados al azar estén conectados

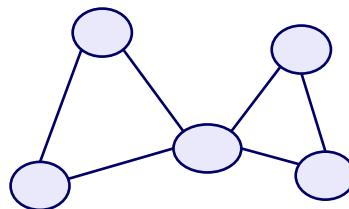
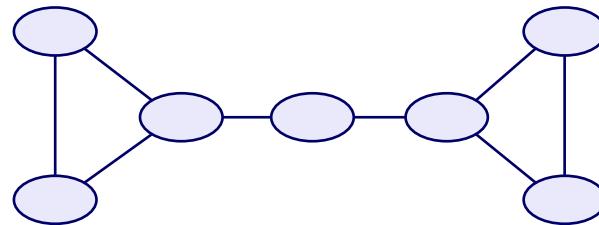
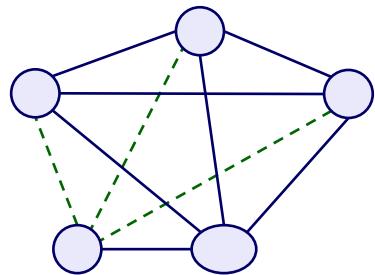
$$C(G) = \frac{\text{nº caminos cerrados de long 2}}{\text{nº de caminos de long 2}} \in [0,1]$$

- ◆ Equivalentemente a la definición anterior, fracción de tripletas transitivas

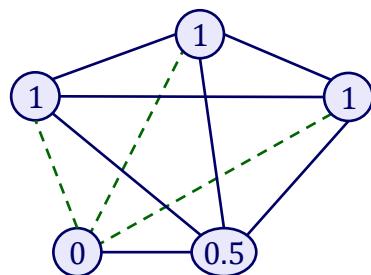
$$C(G) = \frac{3 \times \text{nº triángulos en la red}}{\text{nº de tripletas conectadas}}$$

- ◆ Definición alternativa: $C_{\text{avg}}(G) \equiv \text{avg}_u C(u)$
 - En general se prefieren las definiciones anteriores (en ésta dominan los nodos de bajo grado)
- ◆ El coef de clustering de una red depende de cómo se forman las amistades
 - Si se formasen al azar, sería bastante bajo
 - Mucho más alto si se forman por mediación de amigos, similitud, popularidad...

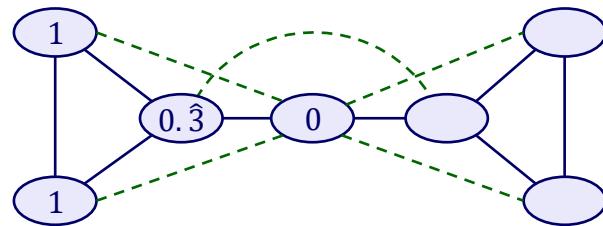
Ejemplos



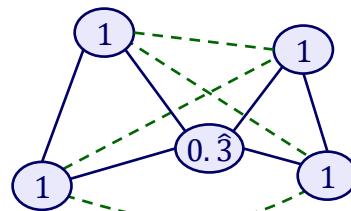
Ejemplos



$$C(G) = \frac{3 \cdot 4}{15} \quad C_{\text{avg}}(G) = \frac{3.5}{5}$$



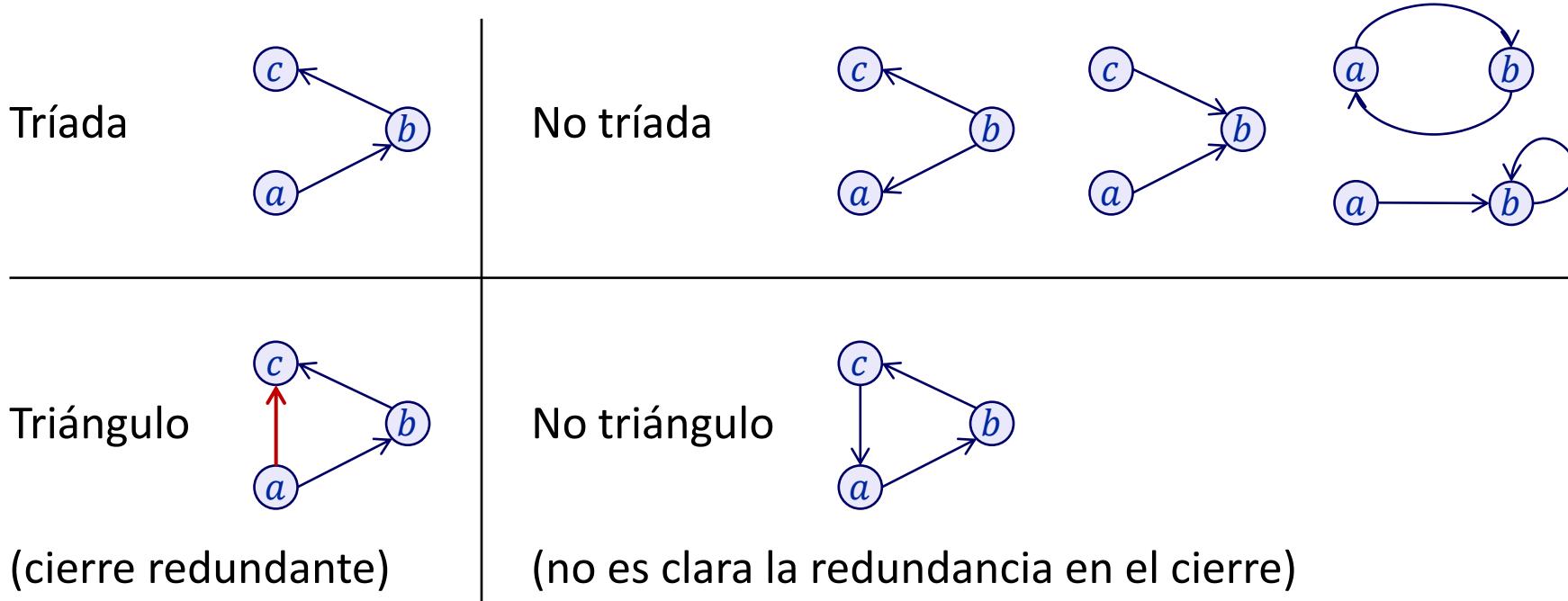
$$C(G) = \frac{3 \cdot 2}{11} \quad C_{\text{avg}}(G) = \frac{4.6}{7}$$



$$C(G) = \frac{3 \cdot 2}{10} \quad C_{\text{avg}}(G) = \frac{4.3}{5}$$

Coeficiente de clustering en redes dirigidas

- ◆ Caben diferentes generalizaciones, una opción común es:
 - Considerar como tríadas las formaciones de tipo $(a, b) + (b, c)$
 - Y como triángulos las formaciones de tipo $(a, b) + (b, c) + (a, c)$
 - Con $a \neq b, a \neq c, b \neq c$

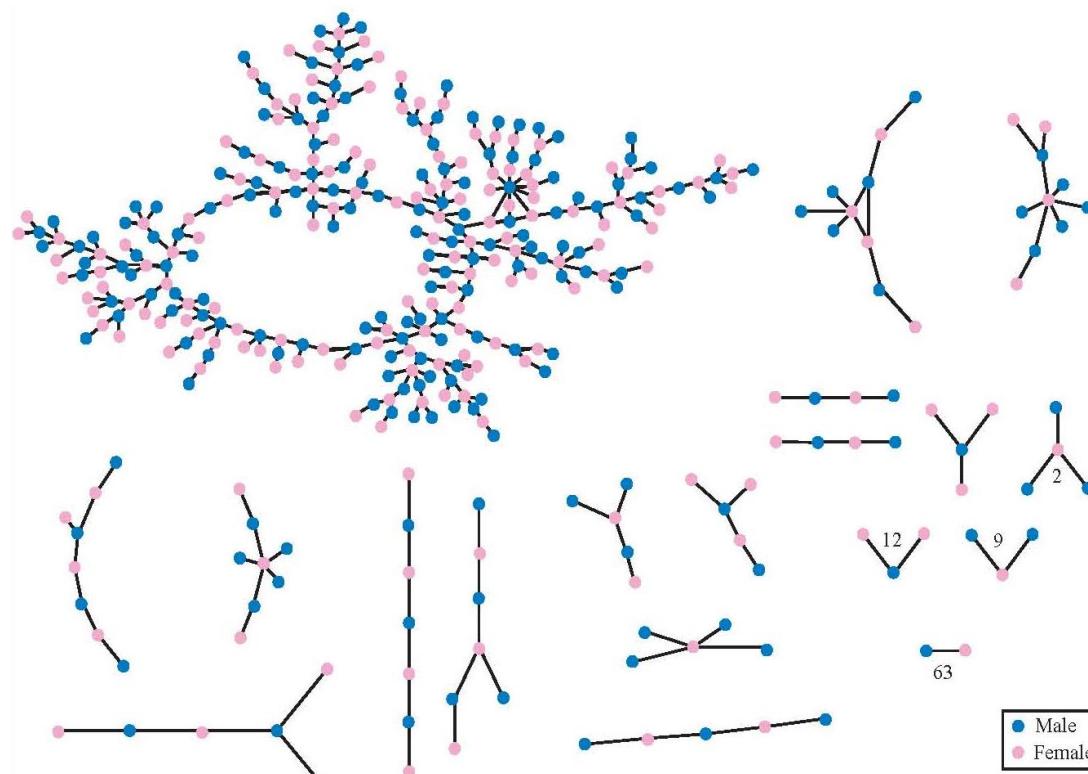


Coeficiente de clustering en redes sociales

- ◆ Habitualmente es “anómalamente” alto comparado con un desarrollo aleatorio de la red: la densidad de las redes sociales tiende a estar fragmentada en comunidades
- ◆ Un amigo común aumenta la oportunidad de enlace
- ◆ Cuando las redes son homófilas, la similitud tiende a menudo a ser transitiva
- ◆ Pueden derivarse ventajas en compartir contactos
- ◆ Existe un factor latente común a la formación de contactos, p.e. la participación en una actividad común (filiación)
- ◆ Y otros posibles factores y teorías...

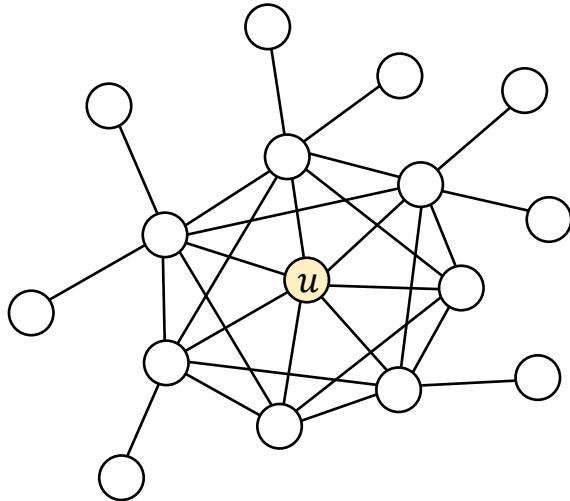
Coeficiente de clustering en redes sociales

- ◆ En ausencia de homofilia...
- ◆ Ejemplo: $CC = 0.005$ (frente a p.e. ~ 0.2 más típico)

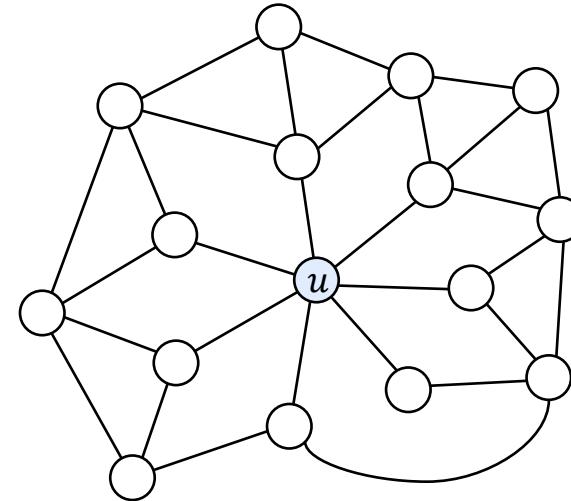


El coeficiente de clustering de una persona refleja cómo de integrada está en un círculo social denso

A



B



También expresa redundancia de relaciones

En buena medida refleja algo inverso a betweenness

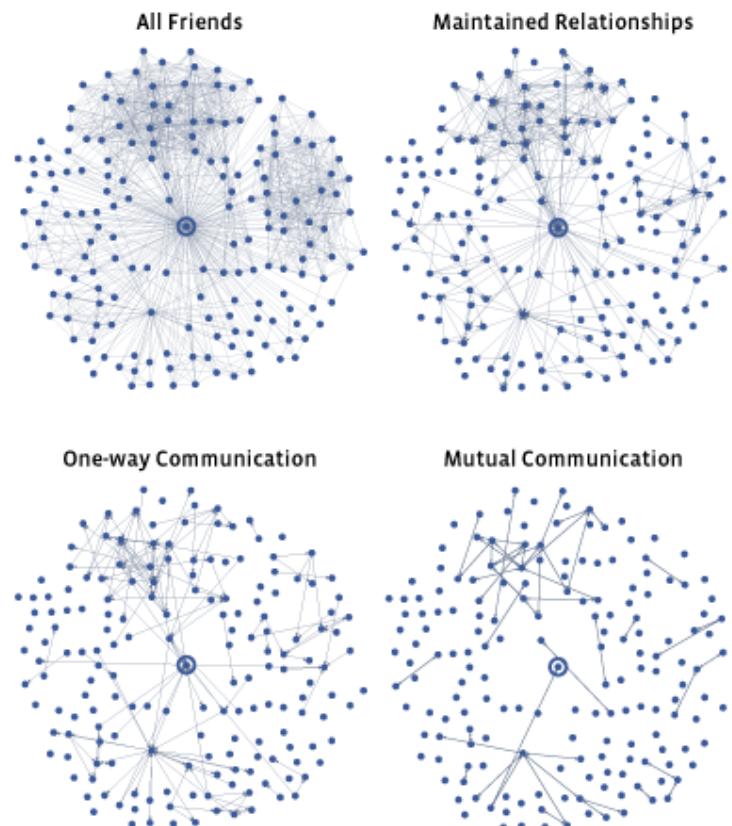
Y para las relaciones, ¿tiene sentido una noción similar?

Métricas sobre enlaces

- ◆ Miden el papel o el valor de un enlace específico entre dos usuarios
 - Tanto si el enlace existe como si no
 - Efecto que aporta para los dos usuarios pero también para la red
- ◆ Muchas medidas giran en torno a nociones de **enlace débil / fuerte**
 - Los enlaces “débiles” suelen tener un valor especial
- ◆ ¿Qué es un enlace débil o fuerte?
 - Definiciones relativas a la interacción: cómo es la relación entre las personas (tipo de relación, frecuencia, duración, semántica, etc.)
 - Definiciones estructurales: cómo es la red en el entorno del enlace (arraigo, puentes, betweenness)

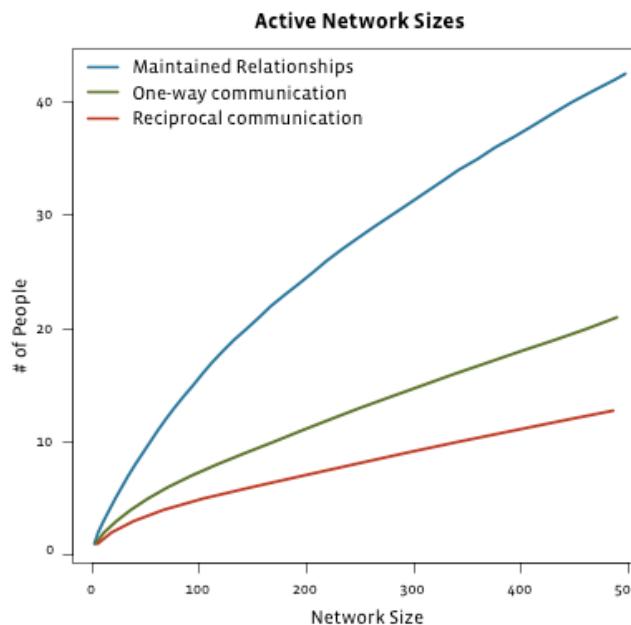
Fuerza de los enlaces: nociones de dominio

- ◆ Nociones de fuerza/debilidad relativas a la interacción y tipo de relación
- ◆ Propias del dominio



[http://overstated.net/2009/03/09/
maintained-relationships-on-facebook](http://overstated.net/2009/03/09/maintained-relationships-on-facebook)

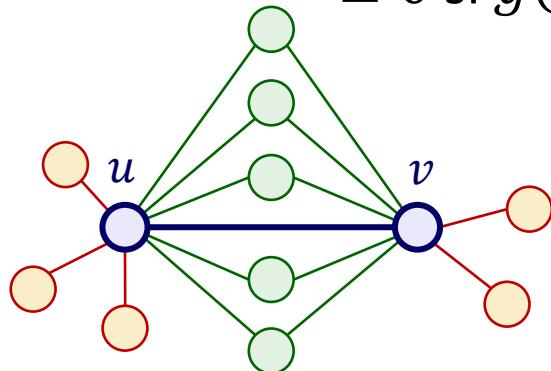
- P.e. simple conexión en una red online vs. frecuencia de interacción directa
- P.e. solidez y estabilidad de la conexión, grados de intensidad, confianza, relación activa vs. pasiva, etc.
- Enlace positivo vs. (implícitamente) negativo
- Correlación: interactuamos más en los entornos con más densidad de enlace (trabajo, etc.)



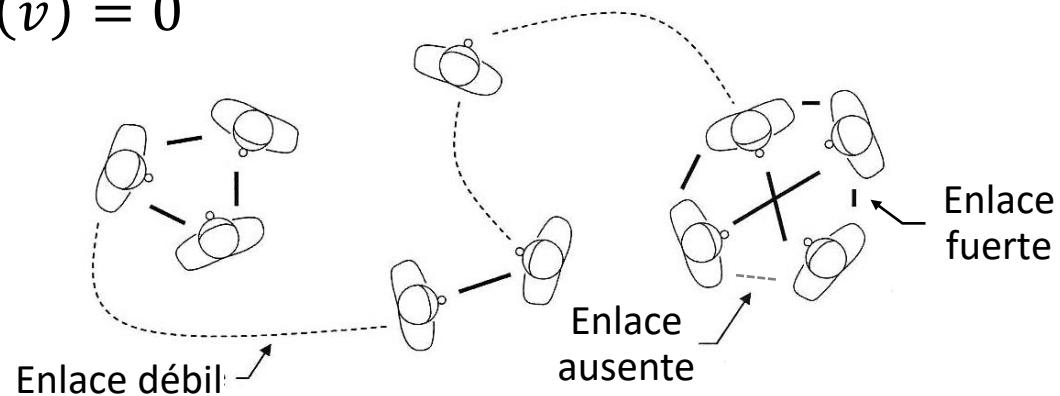
Fuerza estructural: arraigo

- ◆ Solapamiento de vecindarios (embeddedness, overlap...)
- ◆ Arcos arraigados conectan nodos con muchos contactos en común

$$\begin{aligned}\text{Arraigo}(u, v) &= \text{Jaccard}(\text{vecinos}(u) \setminus \{v\}, \text{vecinos}(v) \setminus \{u\}) \\ &\triangleq 1 \text{ si } (u, v) \in A \text{ y } g(u)g(v) = 1 \\ &\triangleq 0 \text{ si } g(u)g(v) = 0\end{aligned}$$



$$\text{P.e. Arraigo}(u, v) = 0.5$$

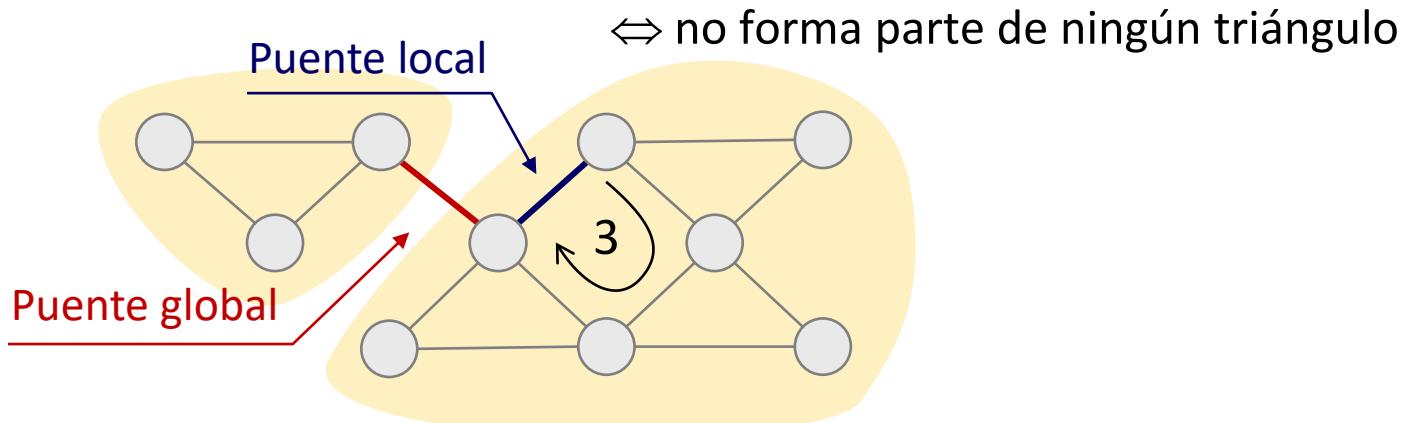


http://en.wikipedia.org/wiki/Interpersonal_ties

- ◆ Es común que el arraigo correlacione con otras nociones de fuerza/debilidad de enlaces propias del dominio

Fuerza estructural: puentes

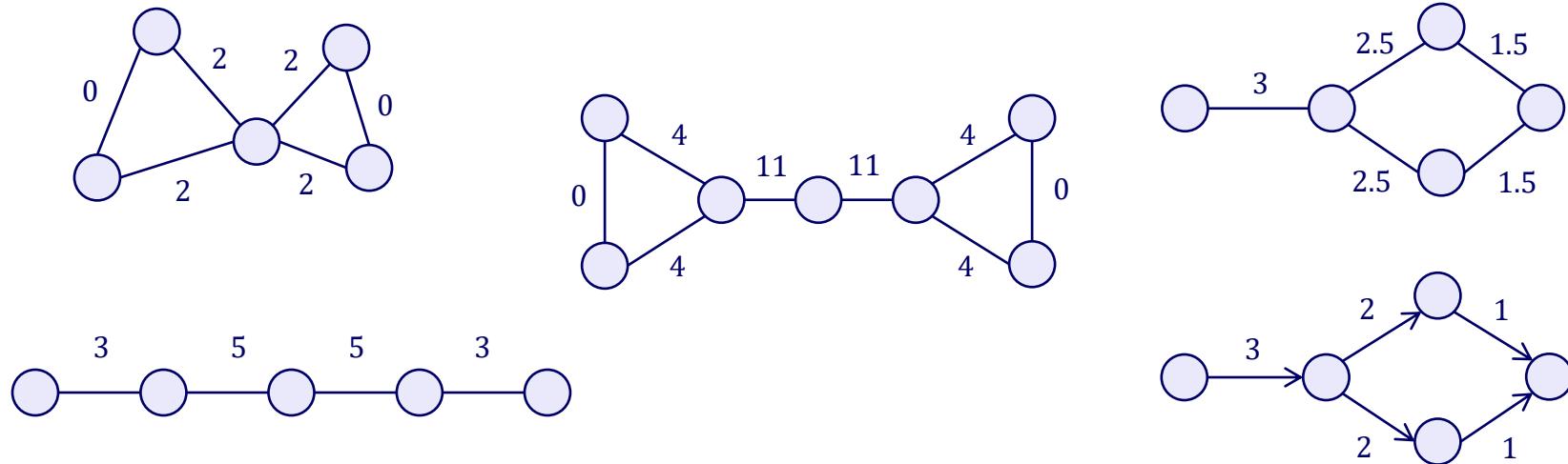
- ◆ Noción relacionada con el arraigo
- ◆ Global: si se elimina el enlace se crea una componente conexa más (cabe considerar puente, puente fuerte –por defecto–, puente débil)
- ◆ Local: si se elimina un enlace (a, b) , entonces $\delta(a, b) > 2$
 - En grafos no dirigidos puente local \Leftrightarrow arraigo 0



- ◆ Punto de vista complementario al coeficiente de clustering local: usuarios con muchos enlaces débiles tienden a bajo clustering

Fuerza estructural: betweenness

- ◆ Se define betweenness de los arcos, igual que de los nodos
- ◆ Promedio (sobre todos los pares de nodos) del ratio de CDMs que pasan por el arco
- ◆ Es decir, para un arco (a, b) , $ns_{v,w}(a, b)$ en lugar de $ns_{v,w}(u)$



Fuerza estructural: interpretación

- ◆ Nodos arraigados vs. mediadores
 - Los enlaces poco arraigados favorecen el alcance de búsqueda y expansión (menos ciclos/redundancias), a pesar de ser débiles (p.e. menos transitados)
 - El comportamiento de los nodos mediadores es clave también: qué harán éstos con sus enlaces débiles, con qué frecuencia los usan, qué grado total tiene el nodo, etc.
- ◆ “Signo” de los enlaces
 - Amigos vs. “enemigos”
 - P.e. ausencia de enlace con fuerte arraigo (i.e. entre personas con muchos contactos comunes) puede indicar aversión

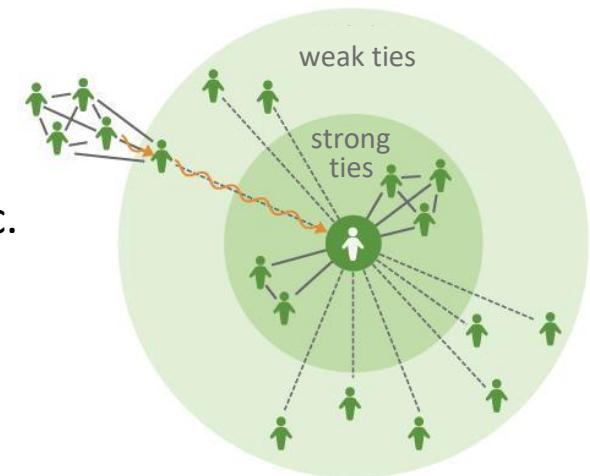
Valor de los enlaces débiles



- ◆ Los enlaces débiles y fuertes tienen distinta utilidad

- Fuertes: ventajas a nivel individual 1-1, p.e. más disponibilidad, fiabilidad, etc. – vivir sin ellos es difícil!
 - Débiles: ventajas en la interacción global, enriquecen y aceleran el flujo de información global, “exclusividad” del contacto a nivel individual, etc.
 - M. S. Granovetter. The strength of weak ties. American Journal of Sociology 78(6), 1973

Mark S. Granovetter
(1943-)



- ◆ Agujeros estructurales

- “People who stand near the holes are at higher risk of having good ideas”
 - R. S. Burt. Structural Holes: The Social Structure of Competition. Harvard University Press, 1995
 - R. S. Burt. Structural Holes and Good Ideas. American Journal of Sociology 110(2), 2004



Ronald S. Burt
(1949-)

Valor de los enlaces débiles (cont)

- ◆ P.e. en Facebook, la mayoría de enlaces son débiles
 - P. De Meo, E. Ferrara, G. Fiumara, A. Provetti. On Facebook, most ties are weak. Communications of the ACM 57(11), 2014
- ◆ Además la mayor parte del flujo de información en Facebook transcurre a través de enlaces débiles
 - E. Bakshy, I. Rosenn, C. Marlow, L. Adamic. The role of social networks in information diffusion. WWW 2012

Relación entre métricas

- ◆ En redes naturales, las siguientes métricas tienden a tener distribución long tail y correlacionan con el grado
 - Betweenness
 - Coeficiente de clustering (inversamente)
 - PageRank
- ◆ Tiene distribución irregular más o menos centrada en la media, y correlaciona menos con el grado
 - Closeness
- ◆ Estas correlaciones pueden variar sensiblemente según los grafos
 - Ver p.e. presentación de P. Boldi en WOA 2012
<http://boldi.di.unimi.it/woa.pdf>

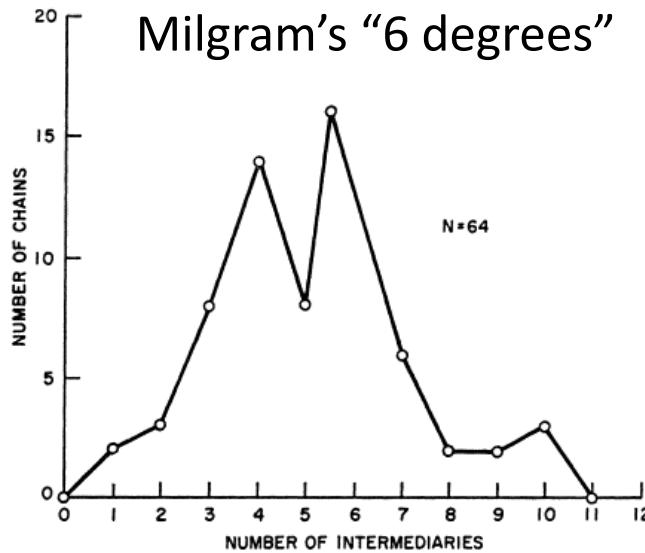
Propiedades globales de las redes

- ◆ Ya vistas:
 - Grado promedio (equivalente a densidad)
 - Distribución de los arcos (i.e. del grado de los nodos)
 - Coeficiente de clustering (i.e. cohesión)
- ◆ Distancias mínimas
- ◆ Estructura de comunidades: componentes conexas, cliques & cores, comunidades
- ◆ Asortatividad: en qué medida los usuarios se relacionan con usuarios similares (homofilia) o diferentes (heterofilia)
 - Similitud de tipo (categórica)
 - Similitud escalar
 - Similitud de grado

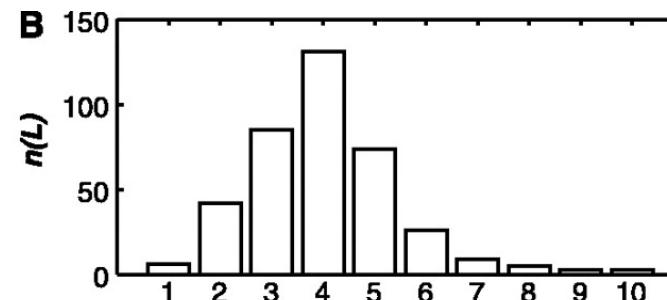
Las redes de gran escala (p.e. $> \sim 100K$ nodos) son difíciles de analizar cualitativamente (p.e. por inspección visual)

Por ello típicamente se observan mediante estas métricas, propiedades y estadísticas globales

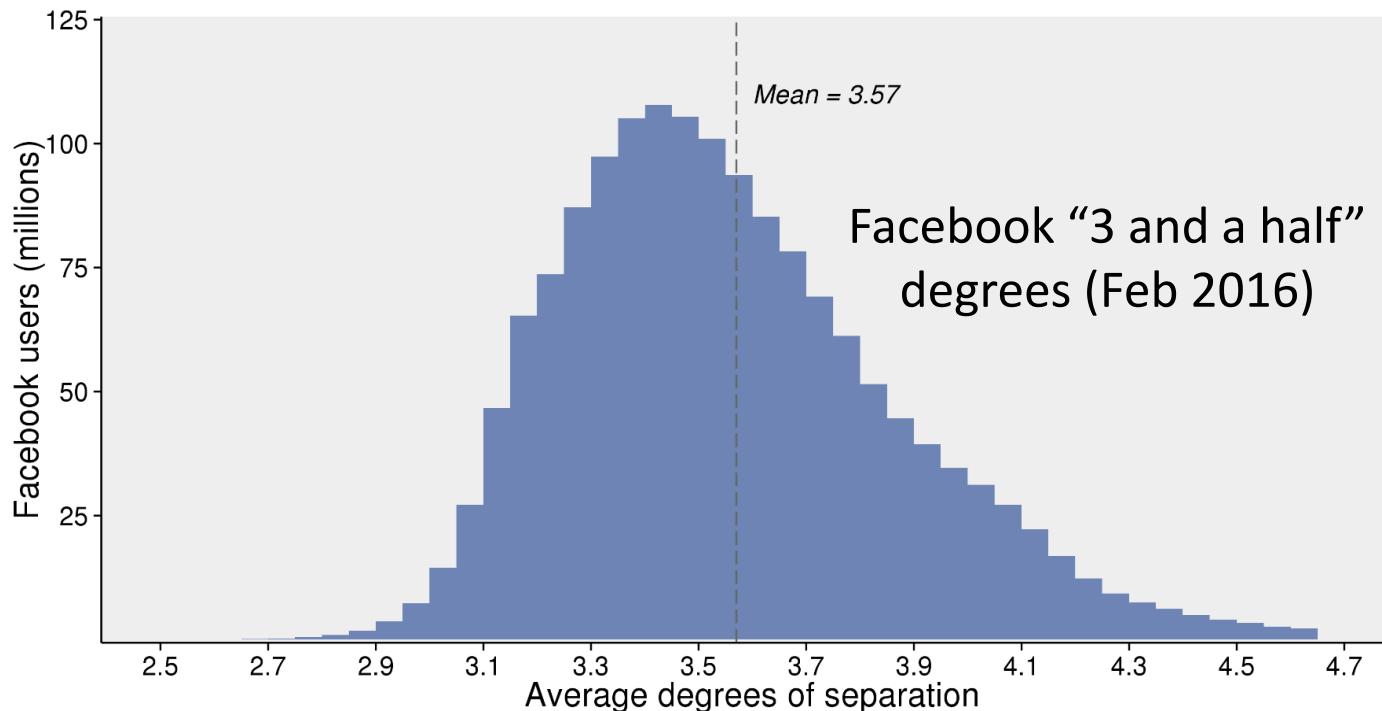
Milgram's "6 degrees"



Distancias mínimas



Dodds et al 2003



Distancias mínimas: promedio

- ◆ “Average shortest path” (ASP)
- ◆ Promedio de las distancias entre todos los pares de nodos

en grafos no dirigidos, 2
si en el sumatorio no se
repiten los pares de nodos

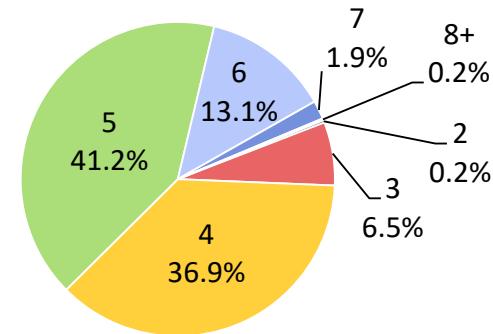
$$\text{ASP} = \frac{1}{n(n - 1)} \sum_{u,v \in V} \delta(u, v)$$

- ◆ En grafos no conexos, se puede medir ASP por separado en cada componente conexa
- ◆ O contar la media restringida a los pares accesibles

$$\text{ASP} = \left(\sum_m \sum_{u,v \in \mathcal{C}_m} \delta(u, v) \right) / \sum_m |\mathcal{C}_m| |\mathcal{C}_m - 1|$$

Distancias mínimas

- ◆ Diámetro: distancia mínima máxima $\max_{u,v} \delta(u, v)$
 - Diámetro exacto, p.e. > 40 en Facebook (estudio de P. Boldi)
 - Diámetro efectivo en percentil 90%: descartando el 10% de diámetros más largos
- ◆ % de usuarios a determinada distancia
 - Promedio por usuario: % de la red accesible a través de k pasos
 - Promedio por pares, p.e. 92% de pares de usuarios a distancia ≤ 5 en Facebook (P. Boldi), 87% en Twitter
- ◆ Etc.
- ◆ Radio: distancia mínima máxima
mínima $\min_u \max_v \delta(u, v)$
- ◆ $\text{radio} \leq \text{diametro} \leq 2 \text{ radio}$



Distribución de distancias en Twitter
<http://www.sysomos.com/insidetwitter>, 2010

Algunos ejemplos de métricas en redes reales

	$ V $	$\text{avg}_u g(u)$	$C(G)$	$C_{\text{avg}}(G)$	ASP	% GC	r	α
Facebook	> 1.000M	140	–	–	4.7	99.9%	–	–
Actores	~450.000	113.4	0.20	0.78	3.48	98%	0.208	2.3
Coautoría math	~250.000	3.92	0.15	0.34	7.57	82.2%	0.120	–
Mensajes email	~60.000	1.44	–	0.16	4.95	95.2%	–	1.5/2
Jefferson High	573	1.66	0.005	0.001	16.01	50%	- 0.029	–
WWW 2000	~200M	7.2	–	–	16.18	91.4%	–	2.1/2.7
Internet 1999	~10.000	5.98	0.035	0.39	3.31	100%	- 0.189	2.5

3. Subdivisión de redes

Subredes

- ◆ Las redes sociales no tienen una cohesión uniforme
- ◆ Suelen observarse regiones
 - Donde la conectividad es más fuerte
 - Donde la interacción es más intensa
- ◆ Diferentes criterios de identificación de subredes
 - Grupos explícitos
 - Componentes conexas (fuerte o débilmente)
 - Cliques y cores
 - Comunidades

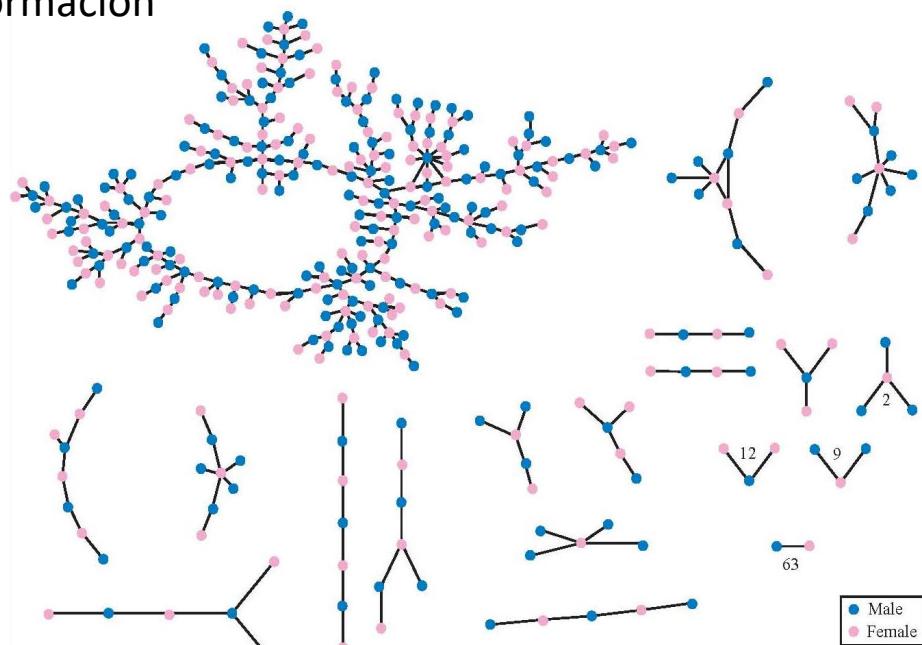
Componentes conexas

- ◆ Subconjuntos de la población que forman una subred conexa maximal – definen una partición
- ◆ Un grafo es conexo \Leftrightarrow sólo tiene una componente conexa
- ◆ Variantes de la definición para redes dirigidas
(para no dirigidas las tres variantes son equivalentes)
 1. **Fuertemente conexa** \Leftrightarrow todo par de personas son mutuamente accesibles
 2. **Conexa** \Leftrightarrow dadas dos personas, al menos una es accesible desde la otra
 3. **Débilmente conexa** \Leftrightarrow fuertemente conexa si ignoramos la dirección

Fuertemente conexa \Rightarrow conexa \Rightarrow débilmente conexa

Componente gigante

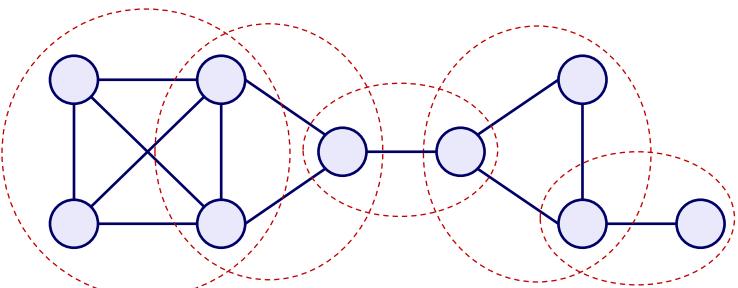
- ◆ Tiende a surgir con mucha facilidad
 - A poco que el grado promedio se acerque a 3
Ver p.e. <https://www.youtube.com/watch?v=HHo50iacrFU>
- ◆ Hay un motivo estadístico: la probabilidad a priori de tener un enlace a la componente conexa más grande es muy alta
 - Es difícil repartir arcos evitando la formación de la componente gigante
 - Salvo que exista un sesgo en otro sentido, p.e. de afinidad categórica
- ◆ Según el tipo de red puede surgir con mayor o menor facilidad
 - P.e. “Jefferson High” componente gigante $> 50\%$ con grado promedio 1.6



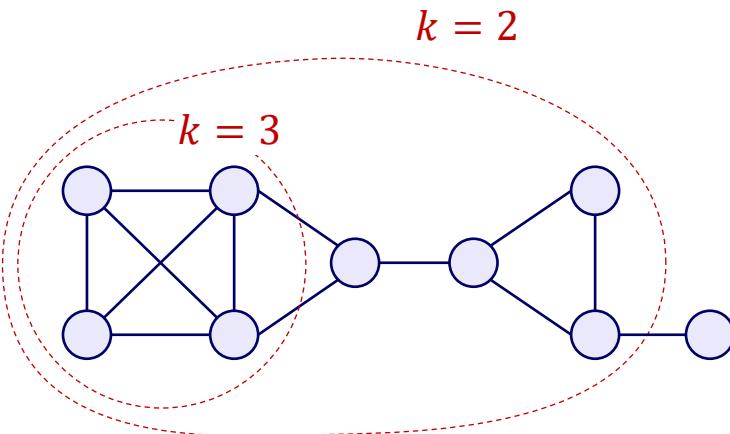
Cliques y otras nociones de cohesión

- ◆ Clique: subgrafo completo maximal
 - Los cliques pueden solapar, i.e. no son una partición del grafo
- ◆ k -core: subgrafo maximal de nodos conectados, al menos, a k otros nodos del subgrafo
 - Los k -cores no solapan
- ◆ δ -clique: subgrafo donde la distancia entre cualquier par es $\leq \delta$
 - Un 1-clique es un clique
 - δ -clan: δ -clique donde la condición se cumple con caminos internos al clique
- ◆ k -plex de tamaño n : subgrafo maximal de n nodos conectados, al menos, a $n - k$ otros nodos del subgrafo
 - P.e. un 1-plex es un clique
 - Un k -plex es un j -plex para todo $j \geq k$
 - Un k -core de tamaño n es un $(n - k)$ -plex
- ◆ Comunidades: subgrafos “densos”

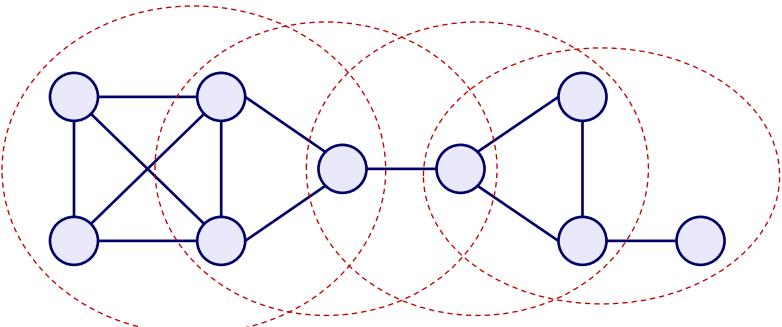
Ejemplos



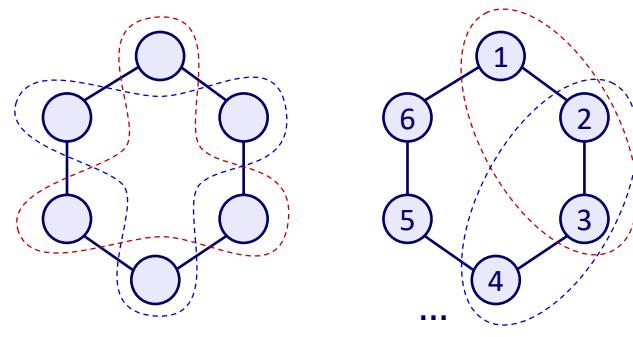
Cliques



k -cores



2-cliques
(todos son 2-cliques)



No clans

2-clans

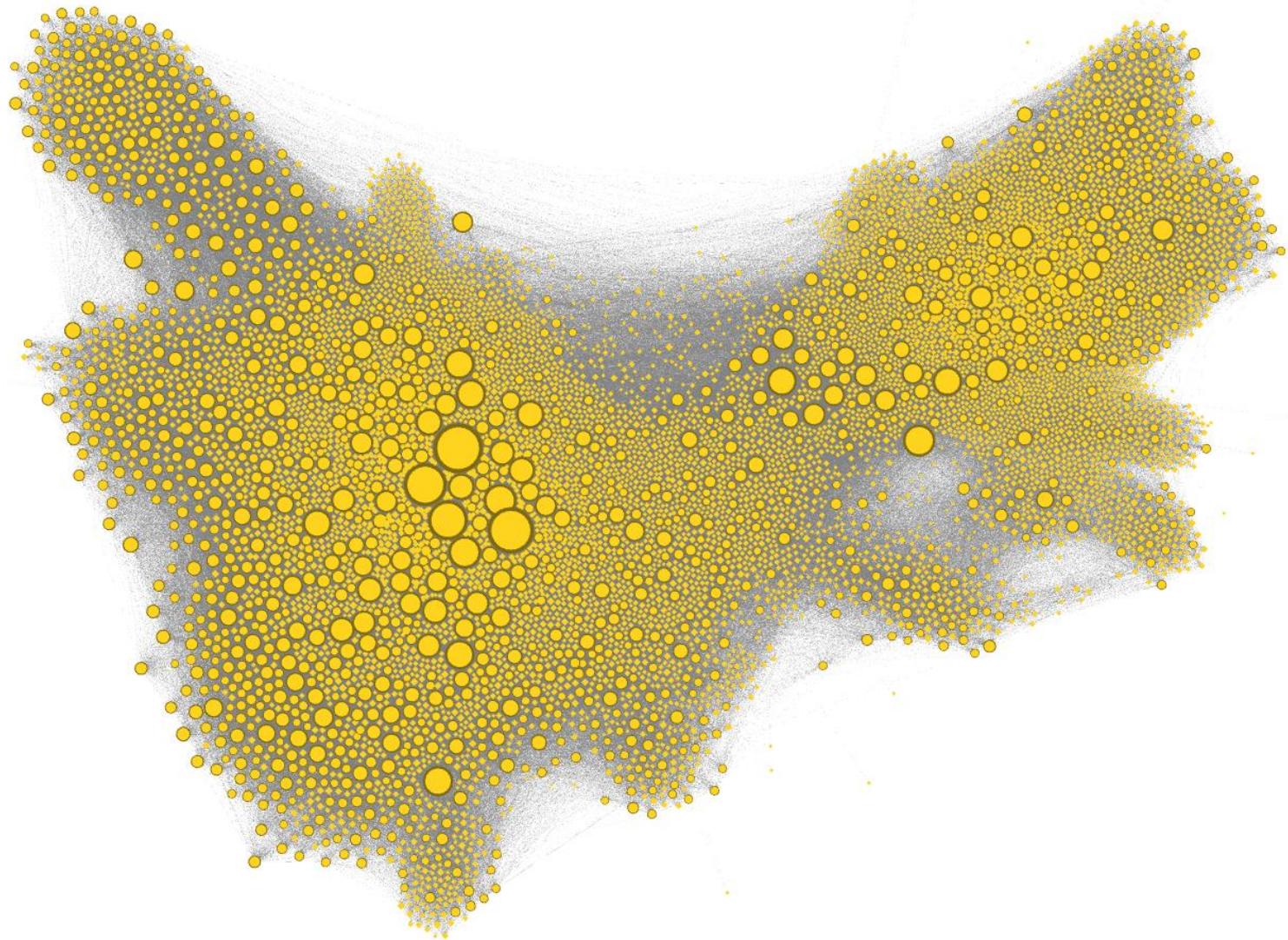
2-cliques

$\{1,2,3\}$
 $\{2,3,4\}$
 $\{3,4,5\}$
 $\{4,5,6\}$
 $\{5,6,1\}$
 $\{6,1,2\}$

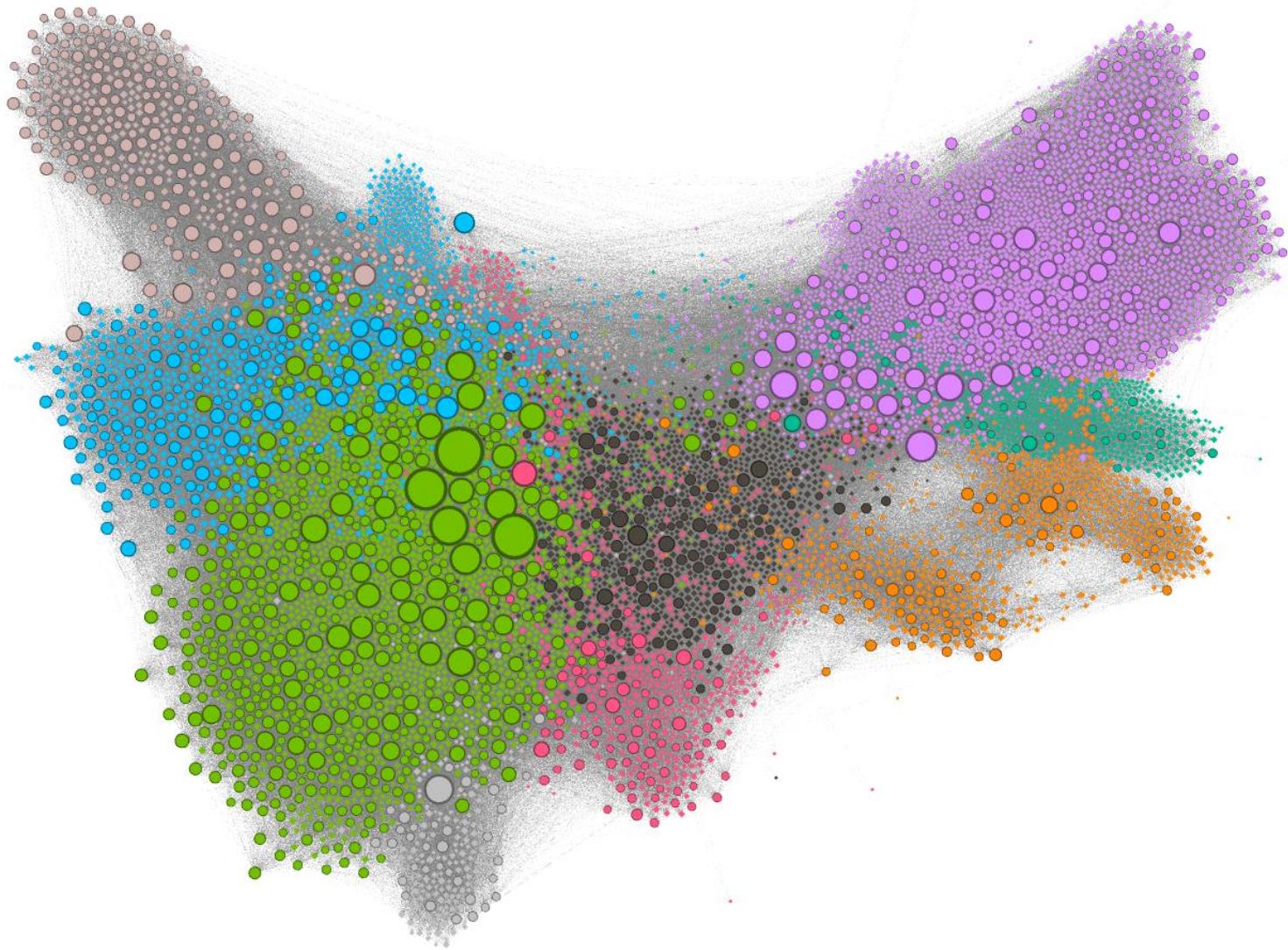
Comunidades

- ◆ Noción relajada de subredes con característica común
- ◆ Pueden ser comunidades explícitas: pertenencia a grupos
- ◆ O pueden detectarse por algún tipo de discontinuidad en la estructura (conectividad, cohesión) de la red
- ◆ Las comunidades explícitas (filiación) suelen tener un reflejo implícito en la estructura de la red
- ◆ Las comunidades implícitas detectables en la estructura de red pueden deberse a grupos (filiación no observada), y/o algún parecido entre las personas que las componen (homofilia)

¿Comunidades?

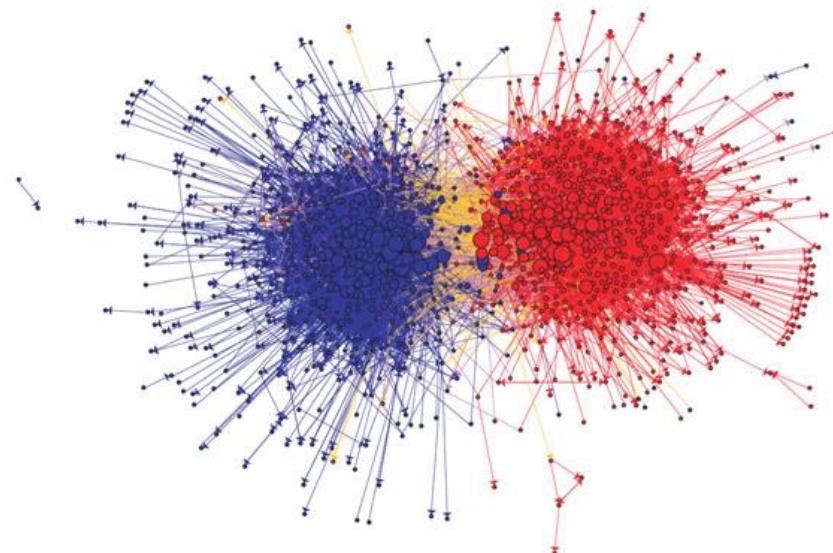


¿Comunidades?



Homofilia

- ♦ Red assortativa: las personas tienden a conectarse con personas similares
 - La assortatividad se define en relación a un cierto criterio de similitud: ubicación geográfica, aficiones, gustos, opiniones, nacionalidad, etnia, ocupación, edad, etc.
 - Efecto bidireccional entre similitud y conexión social: selección vs. influencia
- ♦ Es común observar un cierto grado de assortatividad en las redes naturales
 - Pero se también da el caso contrario (p.e. relaciones de pareja resp. género)



- Conservador
- Liberal
- Cita a post en blog

L. A. Adamic and N. Glance. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. LinkKDD 2005.

Asortatividad: tipos y métricas

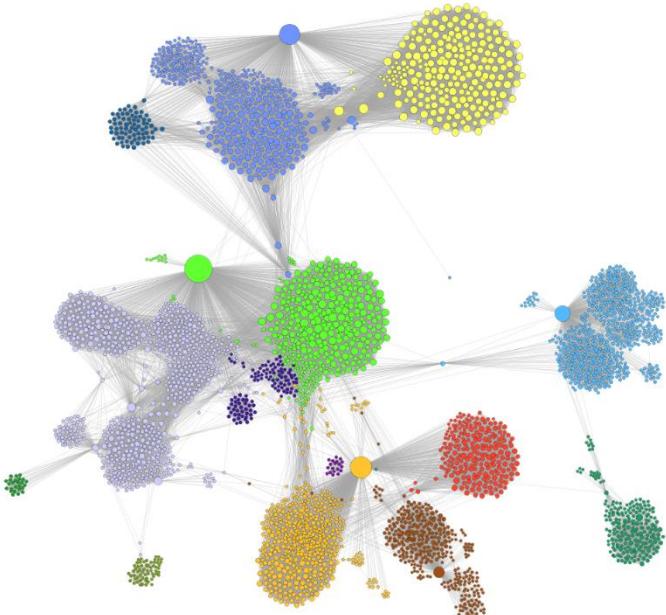
La asortatividad se puede medir respecto a...

- ◆ Categorías → modularidad
- ◆ Un valor escalar → correlación
 - Caso particular: asortatividad de grado
- ◆ Cualquier medida de similitud

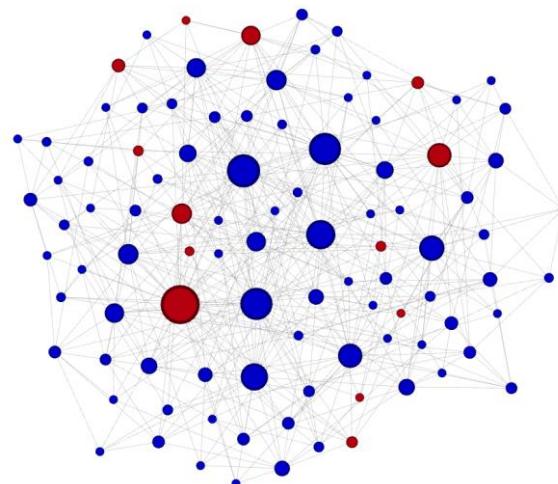
Modularidad

- ◆ Dada una partición de una red (tipología de nodos, grupos)
- ◆ Modularidad = muchos arcos internos a los grupos, pocos arcos entre grupos

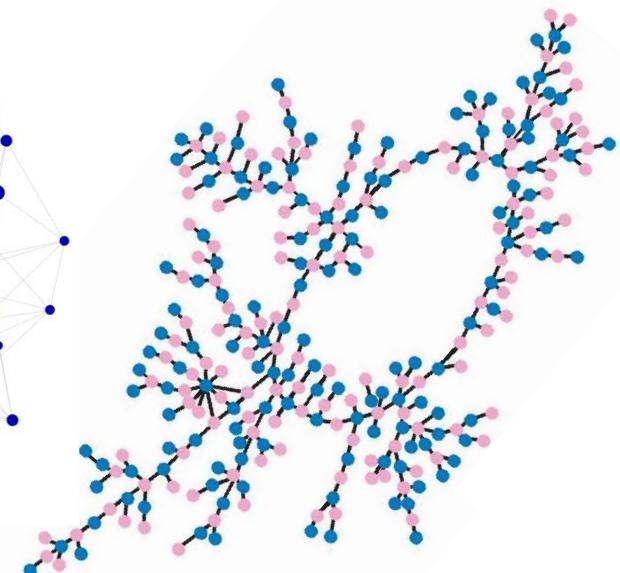
Red muy modular



Red poco modular



Red... ¿anti-modular?



Modularidad

- ◆ Respecto a categorías (tipos, clases, partición)
- ◆ Los nodos son de distintos tipos y la modularidad se mide por el nº de arcos entre nodos del mismo tipo

Probabilidad de que un arco de la red conecte nodos del mismo tipo

Probabilidad de que un arco aleatorio (dados los tipos y grados de la red) conecte nodos del mismo tipo

Nº de arcos entre nodos del mismo tipo

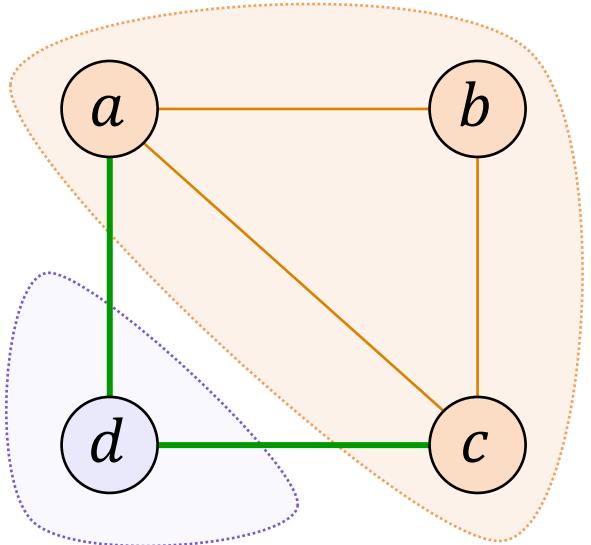
Probabilidad de un arco elegido al azar conecte a u y v

$$r = \frac{\sum_{u \rightarrow v} [c(u) = c(v)]/m - \sum_{u,v} [c(u) = c(v)] g(u)g(v)/4m^2}{1 - \sum_{u,v} [c(u) = c(v)] g(u)g(v)/4m^2} \in [-1,1]$$

(sin repetir arcos)

Máxima diferencia posible del numerador

Modularidad: ejemplo



$$\text{Modularidad} = \frac{\frac{3}{5} - \frac{68}{4 \cdot 25}}{1 - \frac{68}{4 \cdot 25}} = -0.25$$

$$68 = 3 \cdot 3 + 3 \cdot 2 + 3 \cdot 3 + 2 \cdot 2 + 2 \cdot 3 + 2 \cdot 3 + 3 \cdot 3 + 3 \cdot 3 + 3 \cdot 2 + 2 \cdot 2$$

- Hay 3 enlaces internos y 2 externos ¿Cómo puede ser negativa la modularidad?
- Con 3/4 de nodos de la misma clase, entre ellos 2 nodos de alto grado, es fácil tener enlaces internos por azar
- Por ello la diferencia con el número de enlaces internos esperado refleja una modularidad negativa

Correlación

- ◆ Respecto a un valor escalar
- ◆ Cada nodo u tiene un valor u_x (edad, ingresos, etc.),
y se mide la cercanía de valores de los extremos de los arcos
- ◆ Generaliza la modularidad y equivale al coeficiente de Pearson

$$r = \frac{\sum_{u \rightarrow v} u_x v_x / m - \sum_{u,v} u_x v_x g(u)g(v) / 4m^2}{\sum_u u_x^2 g(u) / 2m - \sum_{u,v} u_x v_x g(u)g(v) / 4m^2} \in [-1,1]$$

- ◆ Reescritura computacionalmente más eficiente

(sin repetir arcos)

$$r = \frac{4m \sum_{u \rightarrow v} u_x v_x - (\sum_u u_x g(u))^2}{2m \sum_u u_x^2 g(u) - (\sum_u u_x g(u))^2} \in [-1,1]$$

Asortatividad de grado

- ◆ Resulta de particularizar la correlación escalar tomando el grado de los nodos como valor $u_x \leftarrow g(u)$

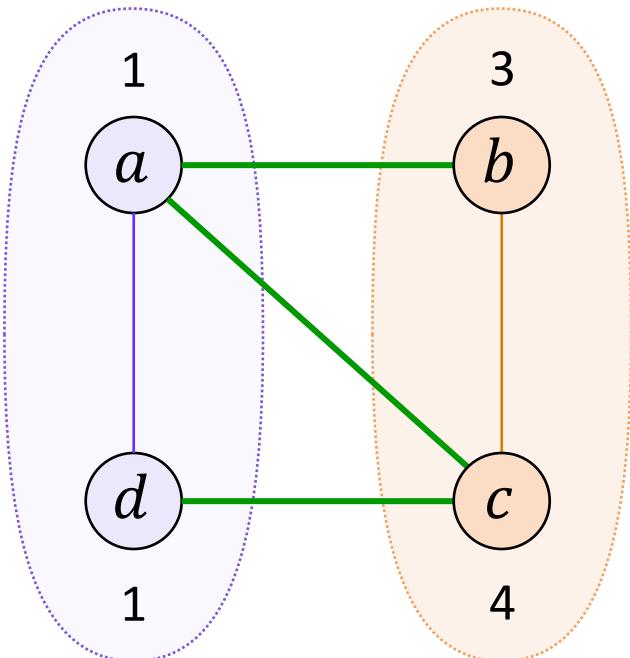
$$r = \frac{\sum_{u \rightarrow v} g(u)g(v)/m - \sum_{u,v} g(u)^2g(v)^2/4m^2}{\sum_u g(u)^3/2m - \sum_{u,v} g(u)^2g(v)^2/4m^2} \in [-1,1]$$

- ◆ Es fácil ver que la fórmula se puede reescribir de forma más eficiente de computar

(sin repetir arcos) 

$$r = \frac{4m \sum_{u \rightarrow v} g(u)g(v) - (\sum_u g(u)^2)^2}{2m \sum_u g(u)^3 - (\sum_u g(u)^2)^2} \in [-1,1]$$

Ejemplo



$$\text{Modularidad} = \frac{\frac{2}{5} - \frac{50}{4 \cdot 25}}{1 - \frac{50}{4 \cdot 25}} = -0.2$$

$$50 = 3 \cdot 3 + 3 \cdot 2 + 2 \cdot 2 + 2 \cdot 3 + 3 \cdot 3 + 3 \cdot 2 + 2 \cdot 2 + 2 \cdot 3$$

$$\text{Correlación} = \frac{4 \cdot 5(1 \cdot 3 + 1 \cdot 4 + 1 \cdot 1 + 3 \cdot 4 + 4 \cdot 1) - 529}{2 \cdot 5(1^2 \cdot 3 + 3^2 \cdot 2 + 4^2 \cdot 3 + 1^2 \cdot 2) - 529} = -0.27$$

$$529 = (1 \cdot 3 + 3 \cdot 2 + 4 \cdot 3 + 1 \cdot 2)^2$$

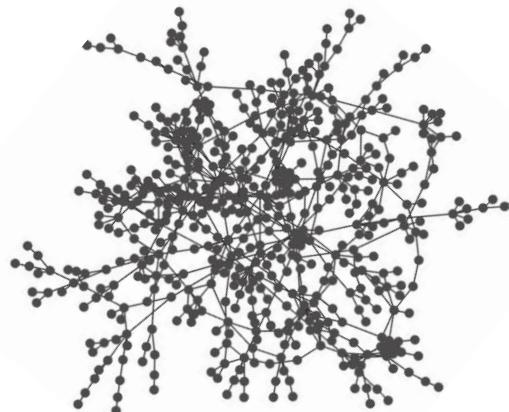
$$\text{Asortatividad de grado} = \frac{20 \cdot 33 - 676}{10 \cdot 70 - 676} = -0.6$$

Asortatividad de grado

- ◆ Las redes sociales tienden a tener un sesgo asortativo $r > 0$
 - En parte se debe a la agrupación en comunidades



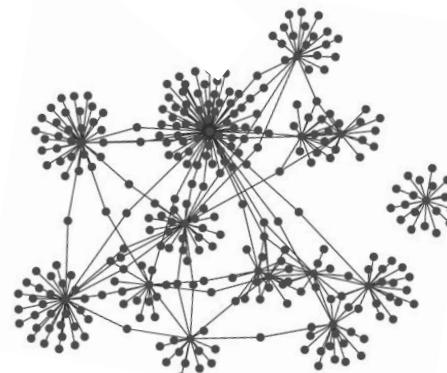
Asortativo



- ◆ Otros tipos de redes (Web, etc.) tienden a $r < 0$
 - En parte se debe a la menor densidad



Desasortativo



- ◆ Erdös-Rényi y Barabási-Albert (ver más adelante) tienen $r \sim 0$

Aplicación a la detección de comunidades

- ◆ Le damos la vuelta al problema
 - En lugar de medir la modularidad dada una subdivisión de nodos en tipos... (p.e. para explicar qué factores influyen en la formación de relaciones)
 - ...dada una red, buscamos una subdivisión que minimice el número de relaciones entre divisiones de la población \Rightarrow que maximice la modularidad
 - Para que así se obtenga una partición en subredes cohesionadas internamente, y poco conectadas externamente
- ◆ Se trata de un problema de clustering con la modularidad como función objetivo
 - Problema NP por el número exponencial del espacio de soluciones (particiones posibles de la población)
- ◆ Soluciones heurísticas muy diversas, típicamente avaras

Algoritmos basados en bisecciones

1. Biseccionar el grafo en dos comunidades aleatoriamente
2. Mover de una comunidad a la otra la persona que más haga aumentar, o menos haga disminuir la modularidad global de la red

Restricción: cada persona sólo se puede mover una vez
3. Repetir 2 hasta haber movido todas las personas
4. Volver atrás y seleccionar el estado de máxima modularidad
5. Se puede volver a 2 con la bisección obtenida, hasta que la modularidad no mejore
6. Volver recursivamente a 1 sobre cada una de las dos comunidades obtenidas

Condición de parada: si una partición no mejora la modularidad, no se hace

$$O(n m \log n)$$

Algoritmos basados en uniones

- ◆ También puede hacerse al revés (p.e. Clauset, Newman & Moore)
 1. Formar comunidades de una sola persona
 2. Unir comunidades de forma avara hasta obtener una sola comunidad
 3. Retroceder y seleccionar el punto donde la modularidad fue máxima

El resultado es algo peor pero es más eficiente $O(n \log^2 n)$ y viable en redes de escala masiva

- ◆ El algoritmo de Louvain (Blondel et al.) se basa en un esquema similar de uniones
- ◆ También puede utilizarse p.e. computación evolutiva o simulated annealing para maximizar la modularidad como función objetivo

Algoritmo de Louvain

Fase 1

1. Crear una comunidad con cada nodo

2. Para cada nodo u

Para cada vecino v de u medir el incremento de modularidad que se produciría si se moviese u a la comunidad de v

Mover u a la comunidad que maximiza este incremento (sólo si el incremento es positivo)

3. Volver a 2 hasta que no se incremente más la modularidad

4. Formar un nodo con cada comunidad

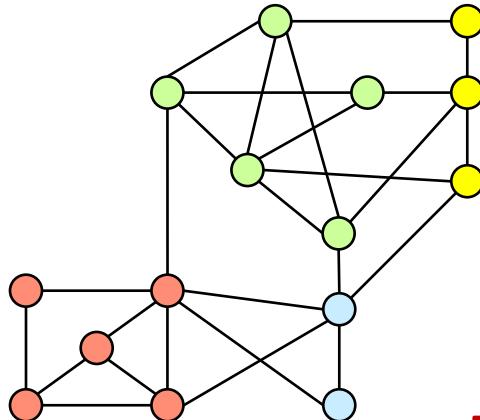
Crear enlaces (y auto-enlaces) entre nodos-comunidad ponderados por el número de enlaces entre los nodos que contienen

3. Volver a 2 con el nuevo grafo hasta que no mejore más la modularidad (ponderada)

Fase 2

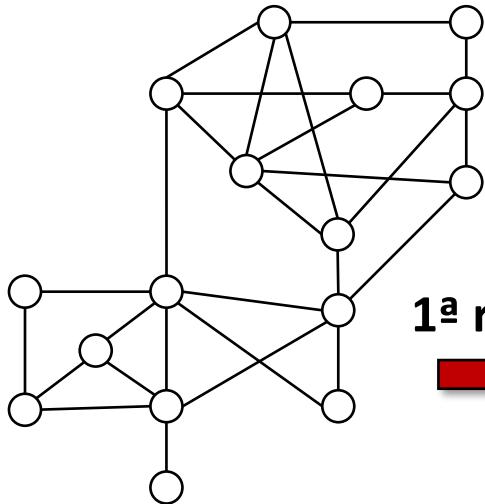
Algoritmo de Louvain

Fase 1
Optimización
de modularidad

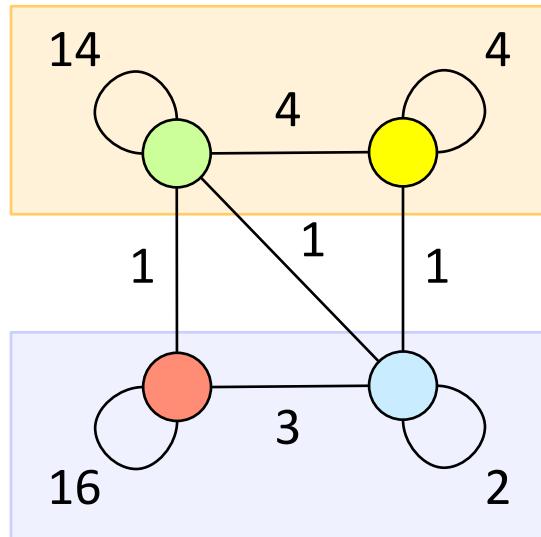


Fase 2

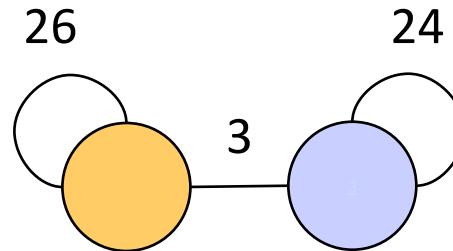
Agregación de
comunidades



1^a ronda



2^a ronda



Algoritmos basados en la eliminación de arcos

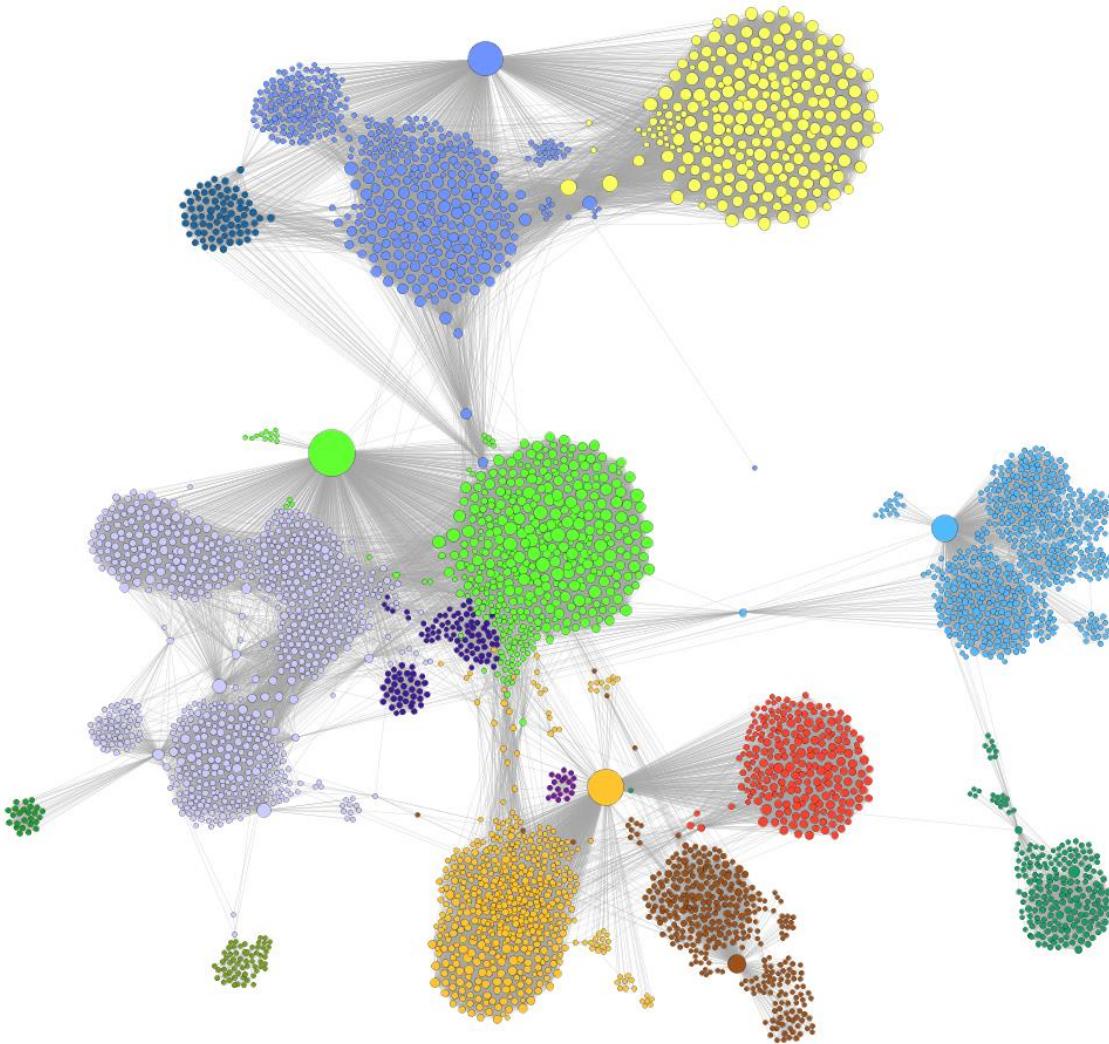
- ◆ Por ejemplo, basado en betweenness (p.e. Girvan-Newman)
 1. Eliminar de la red la relación con mayor betweenness
 2. Recalcular betweenness y volver a 1 hasta eliminar todos los arcos
 3. Resulta un dendograma de componentes conexas (comunidades)
 4. Elegir el nivel que produce p.e. el nº de comunidades deseado, o la máxima modularidad, etc.
- ◆ Nótese que el criterio heurístico no es el mismo que el método basado en modularidad
 - Se basa en que betweenness y modularidad capturan algo relacionado
- ◆ Inconveniente: coste $O(n m(n + m))$, poco viable para $n > 1K$
- ◆ Y muchas otras técnicas más!
 - Basadas en modularidad, betweenness, similitud, u otros criterios...

4. Modelos de formación de redes

¿Qué características presentan típicamente las redes sociales?

- ◆ Nos pueden interesar muy diversas características
 - Distancia promedio, diámetro, pero también...
 - Distribución del grado (sesgada, plana...)
 - Coeficiente de clustering
 - Componente gigante
 - Cómo dependen estas propiedades del tamaño de la red
 - Etc.
- ◆ Problemas
 - Observar y medir las características
 - Explicar y predecir características → **modelos de red**

¿Cómo se forma una red?



Grafo Facebook (J. Leskovec)

[http://snap.stanford.edu/data/
egonets-Facebook.html](http://snap.stanford.edu/data/egonets-Facebook.html)

Red ego 10 usuarios

$$|V| = 4,039$$

$$\text{avg}_{ug}(u) = 43.7$$

$$C_{\text{avg}} = 0.617$$

$$ASP = 3.7$$

$$\text{Diámetro} = 8$$

Modelos de red social

- ◆ Contrastar las redes reales observadas con modelos de formación mucho más sencillos
- ◆ Las diferencias con un modelo aleatorio demostrarían que hay algo no aleatorio en la formación de relaciones en redes reales
- ◆ Los parecidos con un cierto modelo son un indicio de que cierto elemento del modelo se da también en la realidad → pistas hacia una descripción / explicación de las formaciones observadas
- ◆ El modelo más sencillo de todos: ¿cómo sería una red social si se construyese de forma totalmente aleatoria?

Modelos aleatorios de redes sociales

- ◆ ¿Para qué buscar un modelo formal de la estructura y formación de las redes?
- ◆ Es comparable a buscar a qué distribución (gaussiana, Poisson, exponencial, etc.) se parece una muestra de datos
 - Se formula una función de distribución
 - Se ajustan sus parámetros a los datos
 - Y se mide qué tal se ajustan (qué tal se parecen) los datos reales a la distribución teórica
 - A partir de ahí se puede trabajar con la distribución teórica como aproximación a los datos reales
 - Pero en redes sociales no es tan fácil, como veremos...
- ◆ ¿Qué ventaja tiene un modelo analítico?
 - Los modelos teóricos permiten manejos más precisos y económicos
 - Tratamiento analítico, p.e. se pueden “demostrar” propiedades
 - Sirve para hacer simulación de redes “parecidas” al tipo de red con el que queremos trabajar
- ◆ Ello es útil por ejemplo para...
 - Explicar y predecir la aparición de tendencias y patrones estructurales, la respuesta de la red a ciertos procesos, y otras propiedades
 - Razonar sobre hipótesis (“en la medida en que la red se parezca a tal modelo teórico...”)
 - Observar desviación respecto de comportamiento aleatorio

Modelos aleatorios de redes sociales (cont)

- ◆ ¿Qué propiedades estadísticas se busca reproducir con los modelos?
 - Distribución del grado
 - Distancia mínima promedio (ASP)
 - Coeficiente de clustering global
 - Distribución, promedio y correlación mutua de métricas (grado, centralidad, clustering)
 - Componentes (fuerte/débilmente) conexas, componente gigante, tamaños de componente gigante y pequeñas
 - Asortatividad
 - ...
- ◆ En general no es fácil definir un modelo que describa con exactitud las redes sociales reales
 - Debido a la alta complejidad de los procesos reales de formación de las redes
 - Pero se han definido varios modelos que se aproximan en aspectos parciales, y sirven de referencia

Modelos aleatorios de redes sociales (cont)

- ◆ Los modelos se diferencian unos de otros en la forma y probabilidad con la que se eligen los nodos a conectar
 - Algunos modelos son “estáticos” (describen la distribución de arcos que da lugar a la estructura)
 - Otros incluyen el crecimiento del grafo (la adición de nodos) en la descripción del proceso de formación de la red
- ◆ Algunos modelos muy estudiados:
 - Aleatorio Erdös-Rényi
 - Enlace preferente Barabási-Albert
 - Mundo pequeño Watts/Stogatz
 - Amigos comunes, geográficos, encuentro
 - ...

Modelo Erdös-Rényi (1959)

- ◆ “Random attachment”
- ◆ Es el modelo más neutro y aleatorio
- ◆ Cada par de nodos se une con probabilidad p
(no se unen con prob. $1 - p$)

– Dados $u, v \in V$, $p((u, v) \in A) = p$

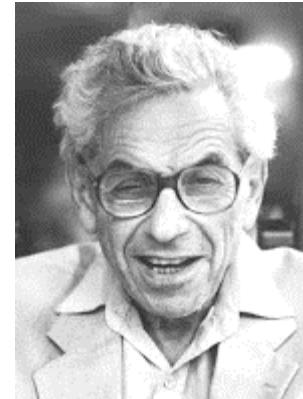
- ◆ Distribución del grado de los nodos: binomial

– $p(g(u) = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$

– Grado promedio = $p(n - 1)$

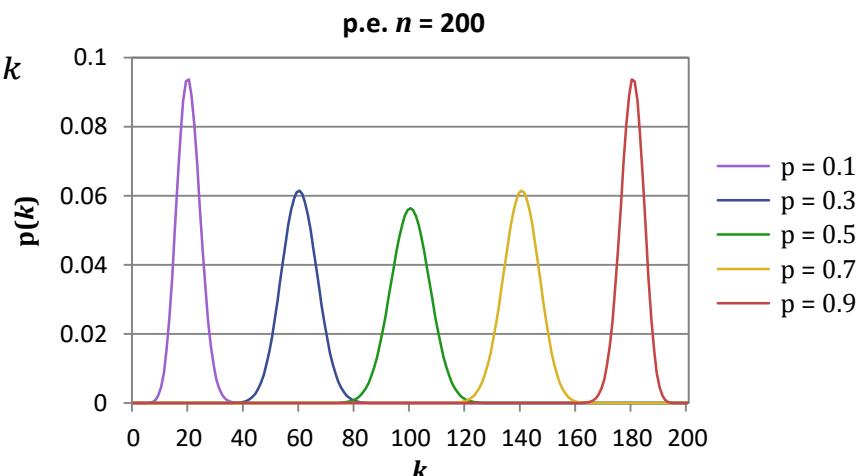
– Nº de arcos esperado = $p \frac{n(n-1)}{2}$

– Para n grande se parece a una Poisson
(por tanto a una gaussiana)

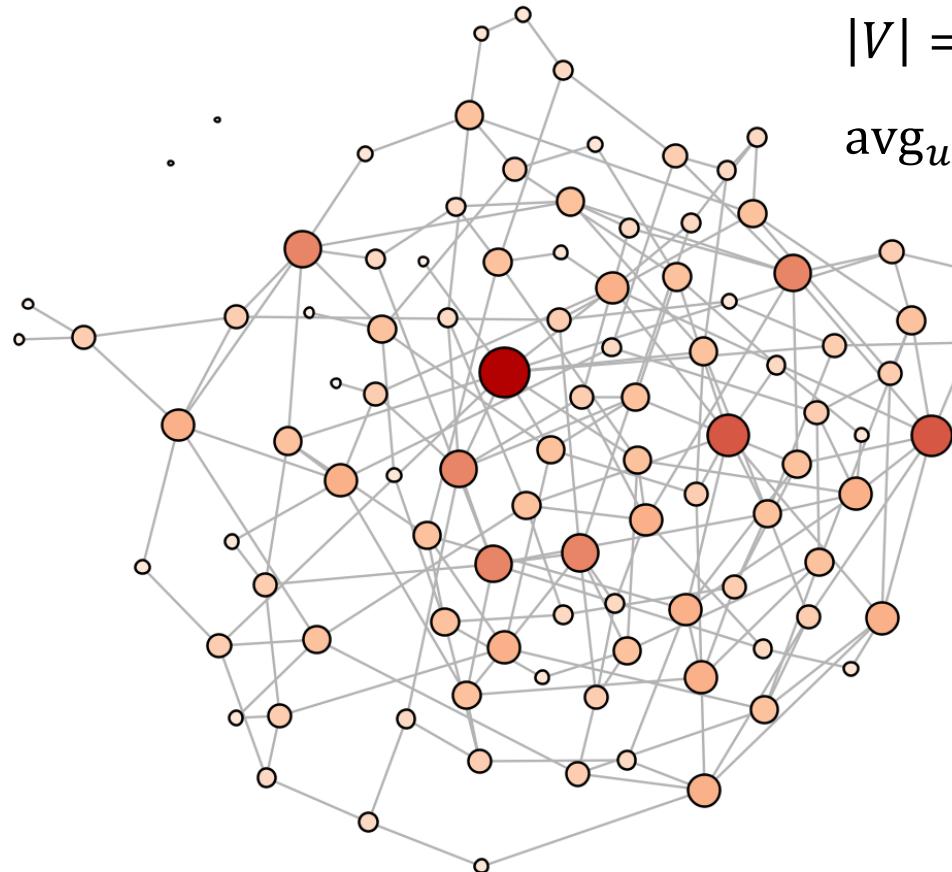


Paul Erdős
(1913-1996)

Alfréd Rényi
(1921-1970)



Ejemplo

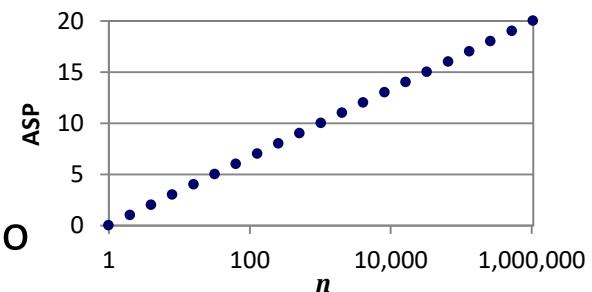
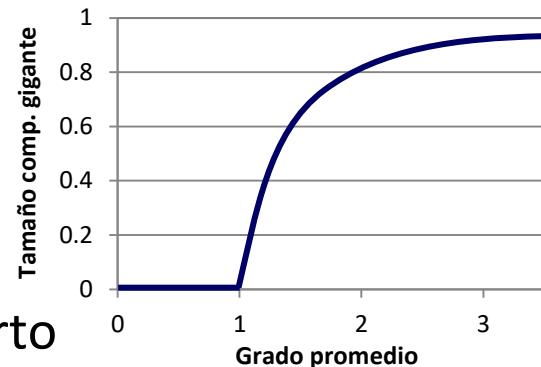


$$|V| = 100$$

$$\text{avg}_u g(u) = 4$$

Modelo Erdös-Rényi (cont)

- ◆ Tiende a generar una componente gigante muy pronto (con $\text{avg}_u g(u) \sim 3$)
 - Ver <https://www.youtube.com/watch?v=HHo50iacrFU>
- ◆ Tiende a producir un camino mínimo promedio corto
 $\text{ASP} \sim \log n / \log \text{avg}_u g(u)$
 - Intuición: árbol de caminos de distancia mínima ASP con ramificación $\text{avg}_u g(u)$ abarcando los n nodos
- ◆ Tiende a generar un coeficiente de clustering muy bajo
$$C(G) = \frac{\text{avg}_u g(u)}{n - 1}$$
 - Esto es muy difícil que se produzca en una red real (social u otras), donde la existencia de un vecino común aumenta la probabilidad de enlace entre dos nodos
- ◆ También es antinatural la ausencia de afinidad por grado, la ausencia de agrupamientos (comunidades), y la distribución binomial sin grandes “hubs”
 - Pero el modelo es útil como punto de referencia y comparación



Modelo de enlace preferente



- ◆ No tan aleatorio, pero también muy simple
 - 1. Los nodos tienen edad: no aparecen todos al mismo tiempo (grafo dinámico)
 - Los nodos antiguos tendrán más enlaces
 - 2. Al crear un enlace los nodos no se eligen con probabilidad uniforme: los que más enlaces tengan son más probables
 - Ventaja acumulativa, “rich gets richer”
- ◆ Las características 1 y 2 se pueden tomar por separado, pero su efecto conjunto es más interesante
 - Entre otras propiedades, surge una distribución power law del grado

Modelo de enlace preferente: construcción

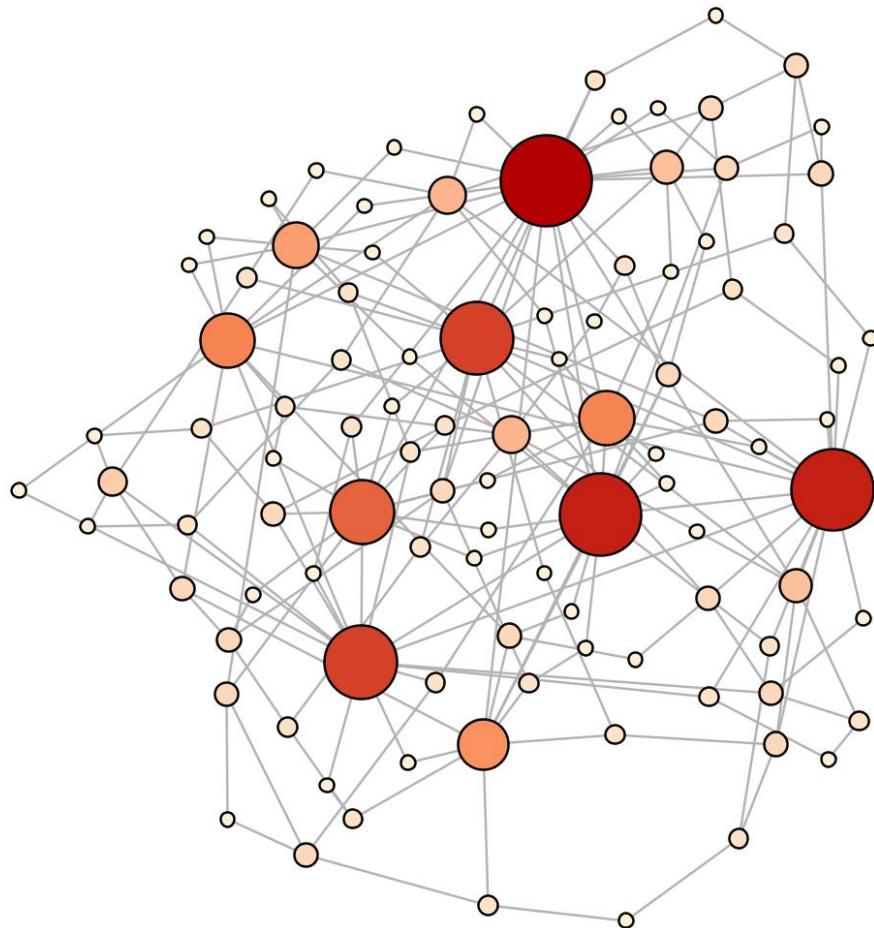
Proceso iterativo de construcción

- ◆ El estado inicial influye poco: p.e. unos pocos nodos enlazados
- ◆ En cada paso se añade un nodo al grafo, y un nº fijo ψ de enlaces de éste a algunos otros elegidos aleatoriamente pero no uniformemente
- ◆ Al introducir un nodo u , la probabilidad de que un nodo v dado sea enlazado por u es proporcional al grado de v

$$p(u \rightarrow v) \sim \frac{\psi g(v)}{\sum_w g(w)}$$

- ◆ Ver <https://www.youtube.com/watch?v=4GDqJVtPEGg>

Ejemplo



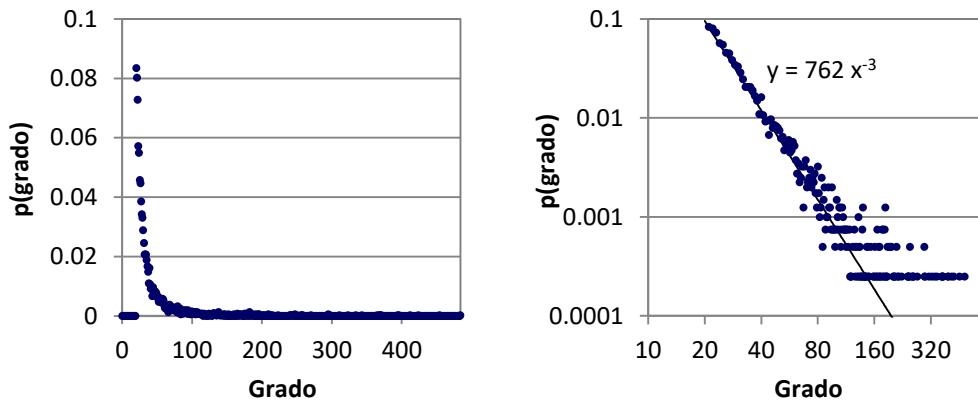
$$|V| = 100$$

$$\text{avg}_u g(u) = 4$$

Modelo de enlace preferente: propiedades

- ◆ Se puede comprobar tanto formalmente como empíricamente que emerge una distribución power law del grado

$$p(g(u) = k) \sim Ck^{-3}$$



- La gráfica del nº de usuarios con cada grado es equivalente (la misma, escalada por n)
- La gráfica de $g(u)$ ordenado está relacionada con ésta

- ◆ ASP $\sim \ln n / \ln \ln n$, y empíricamente $C(G) \sim n^{-0.75}$
- ◆ Efecto colateral: los nodos más antiguos acumulan más enlaces

Modelo de enlace preferente (cont)

- ◆ El proceso es relativamente natural y en alguna medida tiene lugar en fenómenos reales de formación de redes
 - Cita bibliográfica, contactos profesionales, amistades, Web, etc.
 - La popularidad atrae enlaces, por un factor de utilidad o de probabilidad de encuentro
 - De hecho un modelo equivalente al de Barabási emerge escogiendo aleatoriamente un nodo y un vecino de éste al azar (modelo Price)
- ◆ Diferencias con redes reales
 - El **coeficiente de clustering** es menor
 - El **grado crece indefinidamente** con la antigüedad del nodo (se puede ver que es proporcional a la raíz del ratio de antigüedad, excesivo comparado con redes reales)

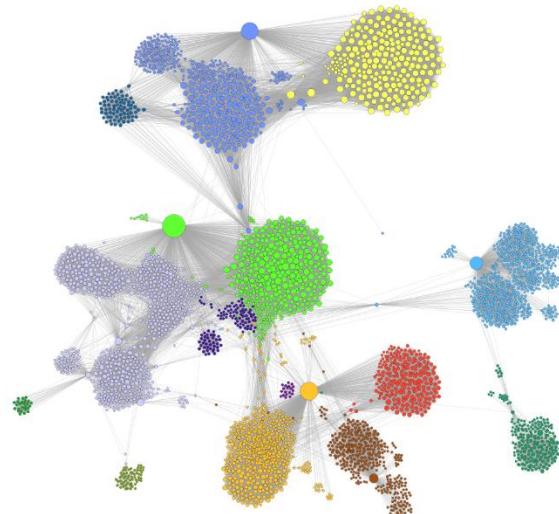
Modelo de enlace preferente (cont)

- ◆ Se han estudiado múltiples generalizaciones del modelo
 - Pueden desaparecer enlaces y nodos (atenúa la ventaja de los usuarios antiguos)
 - La probabilidad de enlace puede ser creciente con el grado de forma no lineal
 - Algunos nodos producen atracción intrínseca independiente de la topología
 - ...con objeto de mejorar el parecido con redes reales
- ◆ El enlace preferente y el crecimiento son necesarios para generar una forma power law
 - El crecimiento con enlace uniforme genera una distribución geométrica
 - El enlace preferente sin crecimiento se aproxima a una gaussiana a medida que se satura la densidad del grafo

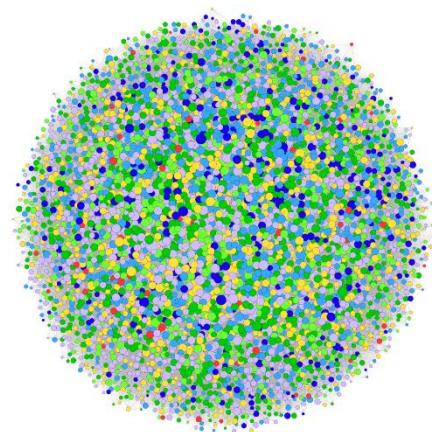
A mayor escala...

$$|V| \sim 4000$$

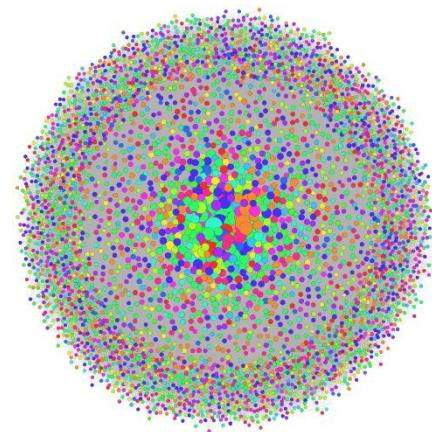
$$\text{avg}_u g(u) \sim 40$$



Facebook 10 ego



Erdős-Rényi

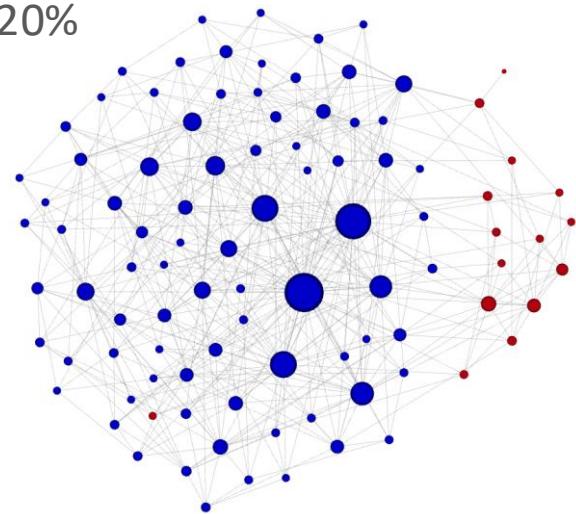


Barabási-Albert

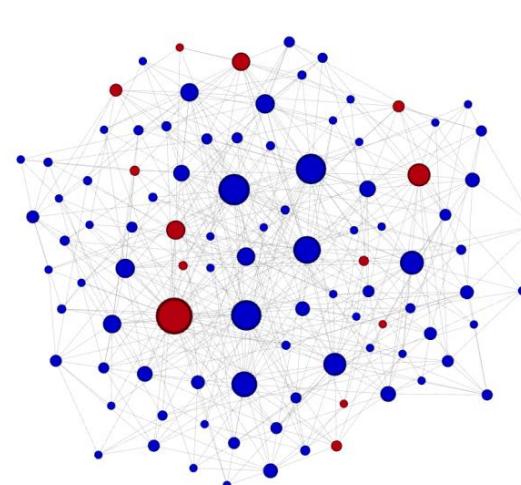
Enlace preferente + homofilia = ?

■ 80%
■ 20%

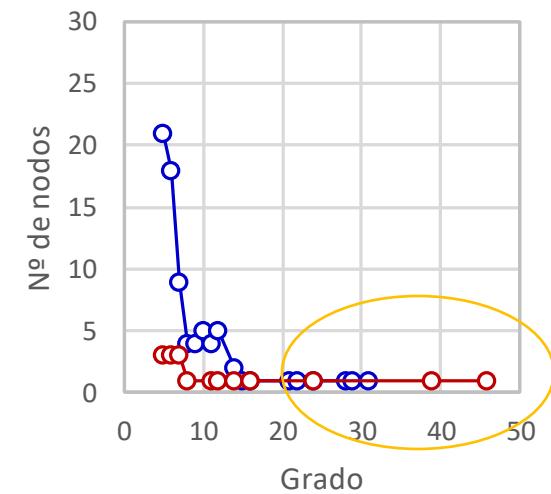
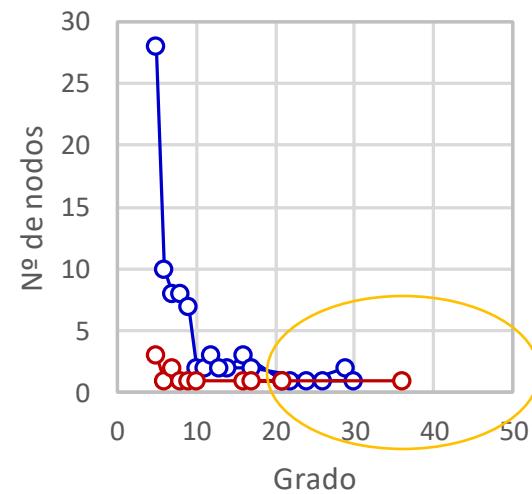
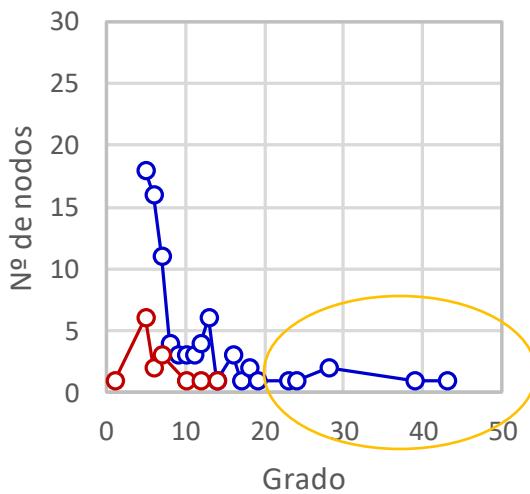
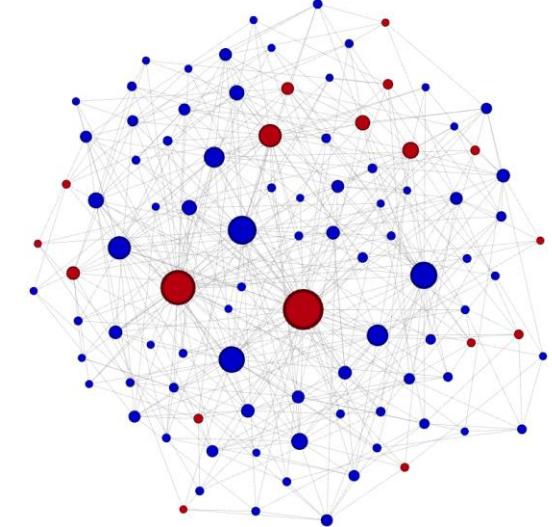
Homofilia



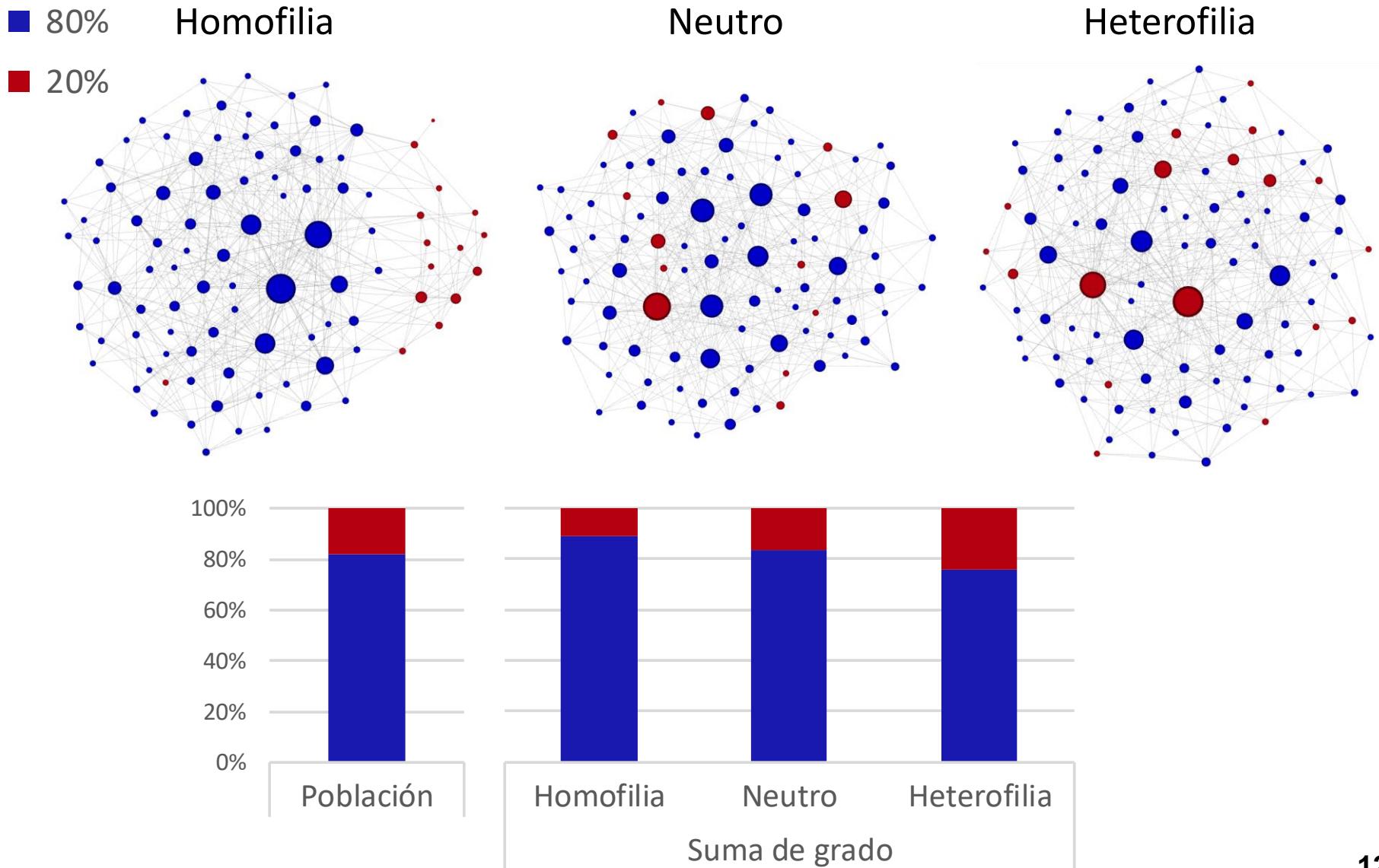
Neutro



Heterofilia

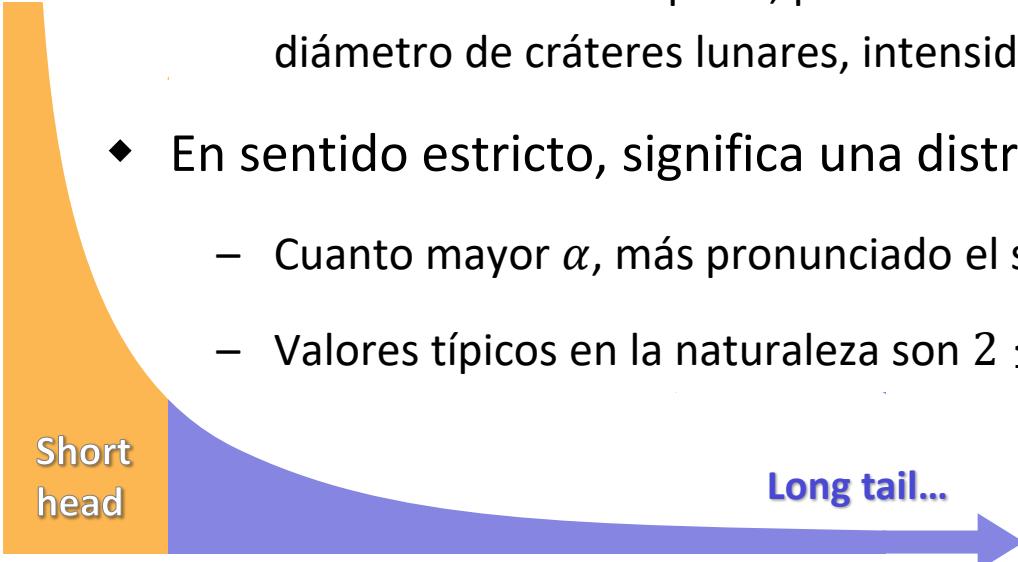


Enlace preferente + homofilia = ?



Fenómenos power law

- ◆ Son recurrentes en fenómenos naturales
 - Distribución de enlaces en redes sociales, de links en la web, frecuencia de las palabras, frecuencia de consultas, nº de clicks, nº de ventas de productos, escuchas de música, nº de ratings de los ítems en sistemas de recomendación, nº de citas de artículos, tamaño de archivos en disco...
 - Distribución de la riqueza, población de las ciudades, frecuencia de apellidos, diámetro de cráteres lunares, intensidad de terremotos, bajas en guerras...
- ◆ En sentido estricto, significa una distribución $p(k) = Ck^{-\alpha}$, con $\alpha > 0$
 - Cuanto mayor α , más pronunciado el sesgo ($\alpha = 0$ es distribución uniforme)
 - Valores típicos en la naturaleza son $2 \leq \alpha \leq 3$



Short head

A diagram illustrating the power-law distribution tail. It features a large orange wedge on the left labeled "Short head". To its right is a blue arrow pointing to the right, labeled "Long tail..." at its tip.

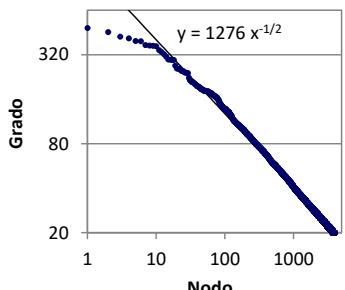
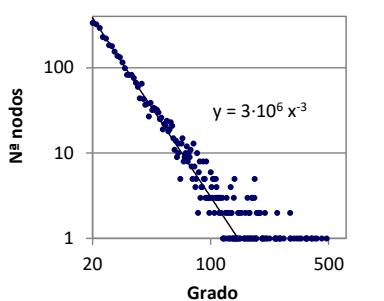
Fenómenos power law (cont)

- ◆ En la práctica rara vez se cumple una distribución power law pura
 - Suele haber desviación en la cabeza y al final de la cola y otras diferencias
 - A menudo debido a condiciones de contorno, p.e. máx 5.000 contactos en Facebook, tratamiento especial de los usuarios nuevos, etc.
 - Se usa no obstante el término “power law” en sentido amplio: acumulación en unos pocos valores (short head), y dispersión en el resto (long tail)
- ◆ Las redes power law se llaman también “libres de escala”
 - Debido a que multiplicando la variable (en este caso el grado) por una constante, la estructura de la distribución no cambia (se escala por la constante elevado a $-\alpha$)

Fenómenos (pseudo) power law y otras distribuciones: ejemplos

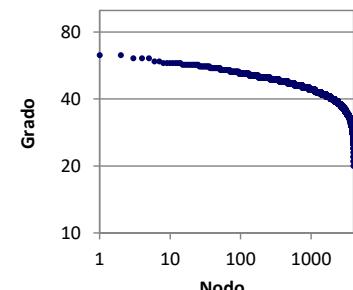
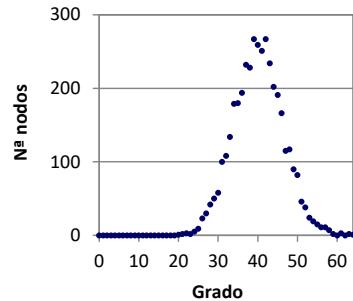
Grafo Barabási-Albert

$|V| = 4,000$
 $\text{avg}_ug(u) = 40$
 $ASP = 2.56 (3.9)$
 $\text{Diámetro} = 4$
 $C_{\text{avg}} = 0.036 (0.002)$



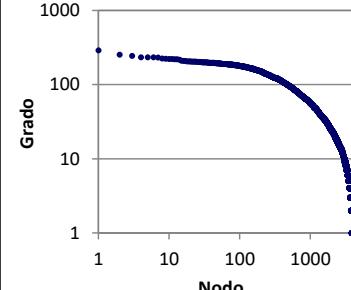
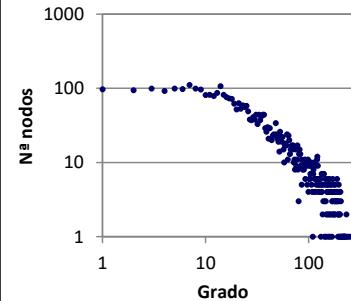
Grafo Erdős-Rényi

$|V| = 4,000$
 $\text{avg}_ug(u) = 40$
 $ASP = 2.65 (2.76)$
 $\text{Diámetro} = 4$
 $C_{\text{avg}} = 0.01 (0.01)$



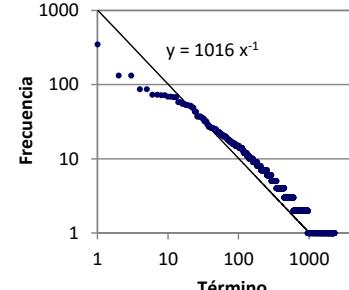
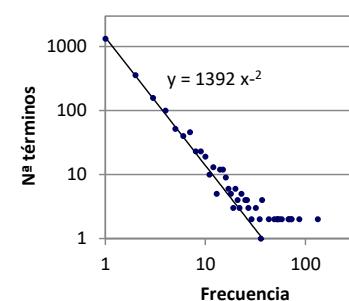
Grafo Facebook (J. Leskovec)

Red ego 10 usuarios
 $|V| = 4,039$
 $\text{avg}_ug(u) = 43.7$
 $ASP = 3.7$ $\text{Diámetro} = 8$
 $C_{\text{avg}} = 0.617$



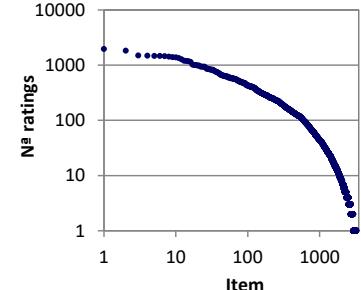
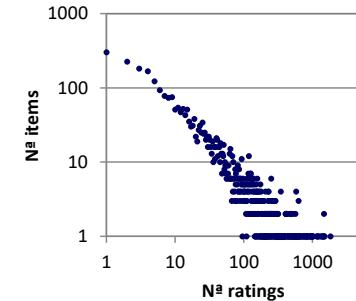
Distribución palabras

$d \equiv \text{http://en.wikipedia.org/wiki/Entropy}$
 $|d| = 8,790$ tokens
 $|\mathcal{V}| = 2,289$ términos



MovieLens 1M

$|\mathcal{I}| = 3,700$
 $|\mathcal{U}| = 6,040$
 $|\{r(u,i) = 5\}| = 226K$



Generados por modelo

Extraídos de datos reales

Distribución en ránking vs. distribución de frecuencias

- ◆ En ciertas estadísticas como la frecuencia de palabras, la propiedad ya es una frecuencia
 - Da pie directamente a un análisis de distribución
- ◆ En otras, como el grado, la propiedad no es necesariamente una frecuencia (o no es interpretada como tal)
 - Se puede visualizar el ránking de los elementos por la propiedad
 - No obstante, el grado también se podría interpretar como la frecuencia con la que un nodo es enlazado por los demás...
- ◆ En este segundo caso es común estudiar la distribución de frecuencias de ese valor (cuando es discreto)
 - Si una distribución es power law, la otra también lo es
 - Los exponentes están relacionados: no es muy difícil comprobar que $\alpha = 1 + 1/\beta$, siendo α el exponente de frecuencia y β del ránking

Redes “de mundo pequeño”

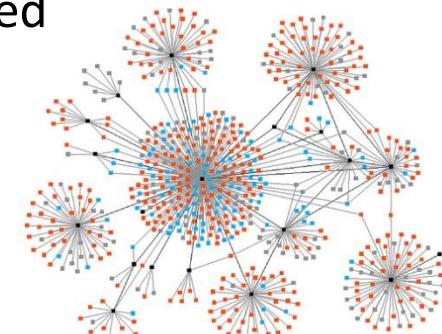
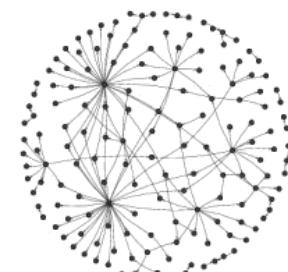
- ◆ Se definen por

- Muy baja probabilidad de que dos nodos al azar estén conectados
- Muy alta probabilidad de que dos contactos de un mismo nodo estén conectados entre sí
- Muy cortas distancias entre nodos aleatorios

- ◆ En términos de métricas

- CDMs sorprendentemente cortos
- ASP relativamente independiente del tamaño de la red
- Distribución power law del grado, grandes hubs
- Coeficiente de clustering muy alto \gg los aleatorios, y relativamente independiente del tamaño de la red

- ◆ El patrón se ha observado sistemáticamente en muy diversas redes sociales y de otros muy diversos tipos



Modelos de mundo pequeño

- ◆ Modelos para generar/explicar redes de mundo pequeño
- ◆ Barabási-Albert no se aproxima del todo (bajo clustering)
- ◆ Se han propuesto otros modelos basados en añadir o reasignar arcos

- Watts/Strogatz 1998, reasignar fracción p de arcos al azar
- G inicial anillo con k vecinos por nodo
- Coef de clustering $C(G_p) \sim C(G)(1 - p)^3$
 $\xrightarrow[k \rightarrow \infty]{}$ $3/4(1 - p)^3$ decrece lentamente con p
- Mientras ASP decrece rápido $\sim \ln(nkp)/k^2p$

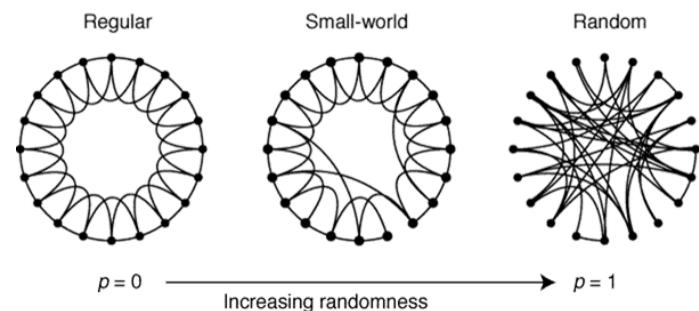
- ◆ Y multitud de variaciones...
- ◆ Teorías que intentan explicar la formación de redes de mundo pequeño
 - Redes de filiación jerárquicas (Kleinberg):
 - Unidades geográficas, estructuras corporativas, etc.
 - Modelos de optimización de coste vs. distancia
 - P.e. red de carreteras vs. rutas aéreas
 - ...



Duncan J. Watts



Steven Strogatz



Modelo “amigos comunes”

- ◆ Los enlaces se forman por mediación de un vecino común (que “presenta” una persona a la otra)
- ◆ Igual que Erdös-Rényi, pero con probabilidad q uno de los nodos se elige entre los vecinos de los vecinos
- ◆ Comparado con Erdös-Rényi, este modelo da lugar a...
 - Coeficiente de clustering mayor
 - ASP más largo
 - Distribución del grado más sesgada
 - Componente gigante más pequeña (para p pequeño)

Modelo geográfico

- ◆ Se define una medida de distancia entre usuarios (p.e. típicamente geográfica)
- ◆ Cada nodo se conecta a sus k vecinos más cercanos
- ◆ Para simular el modelo, se asignan posiciones aleatorias a los usuarios
- ◆ Comparado con Erdös-Rényi, produce ASP más largo

Ejemplo: efecto geográfico en Facebook



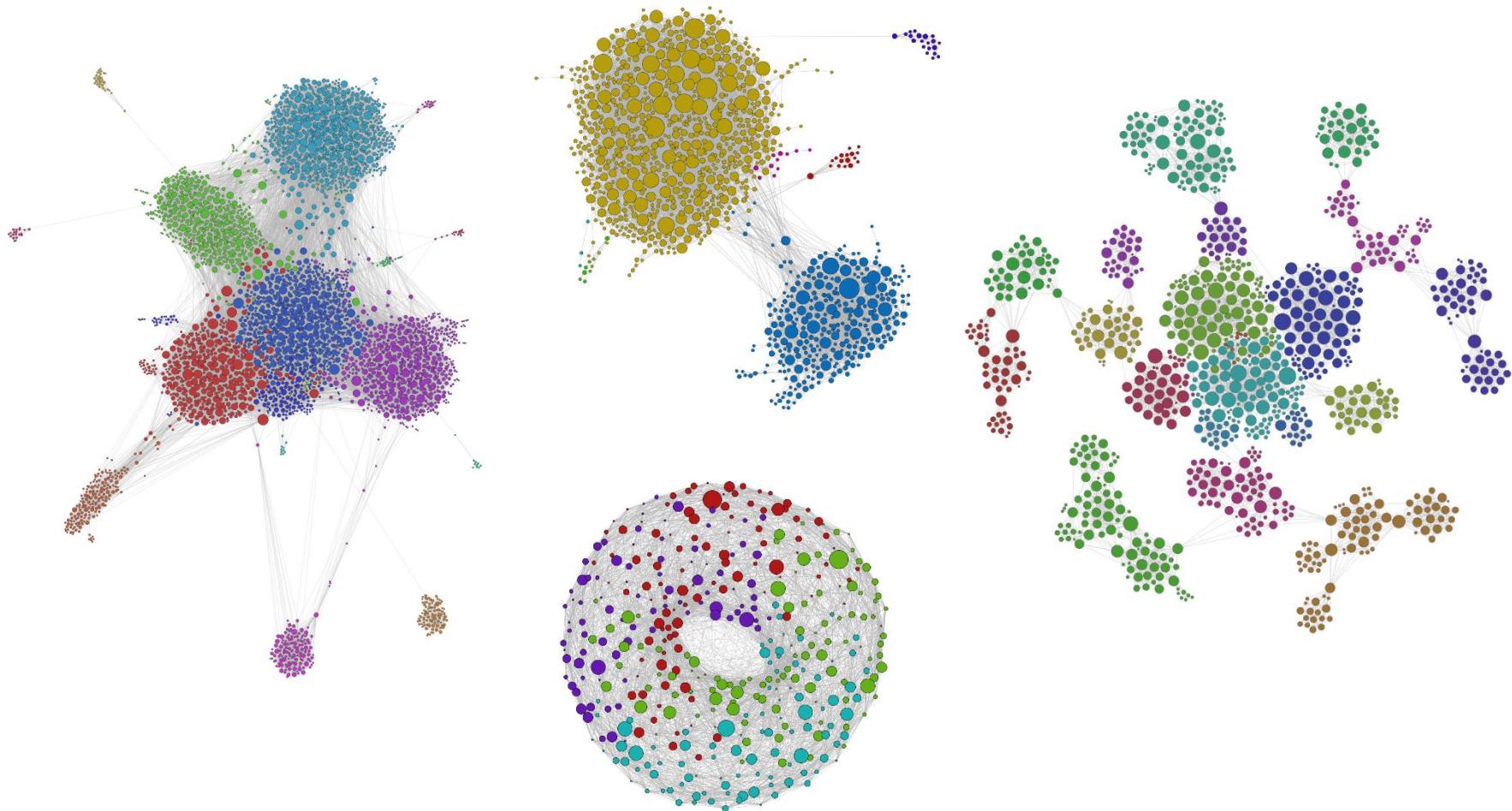
Muestreo de 10M enlaces entre usuarios (2010)

<https://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919>

Encuentro fortuito

- ◆ Los usuarios se mueven en el espacio
- ◆ Cuando se topan unos con otros, se crea un enlace con probabilidad p
- ◆ Comparado con Erdös-Rényi, este modelo da lugar a...
 - Coeficiente de clustering mayor
 - ASP más largo

Otros modelos...



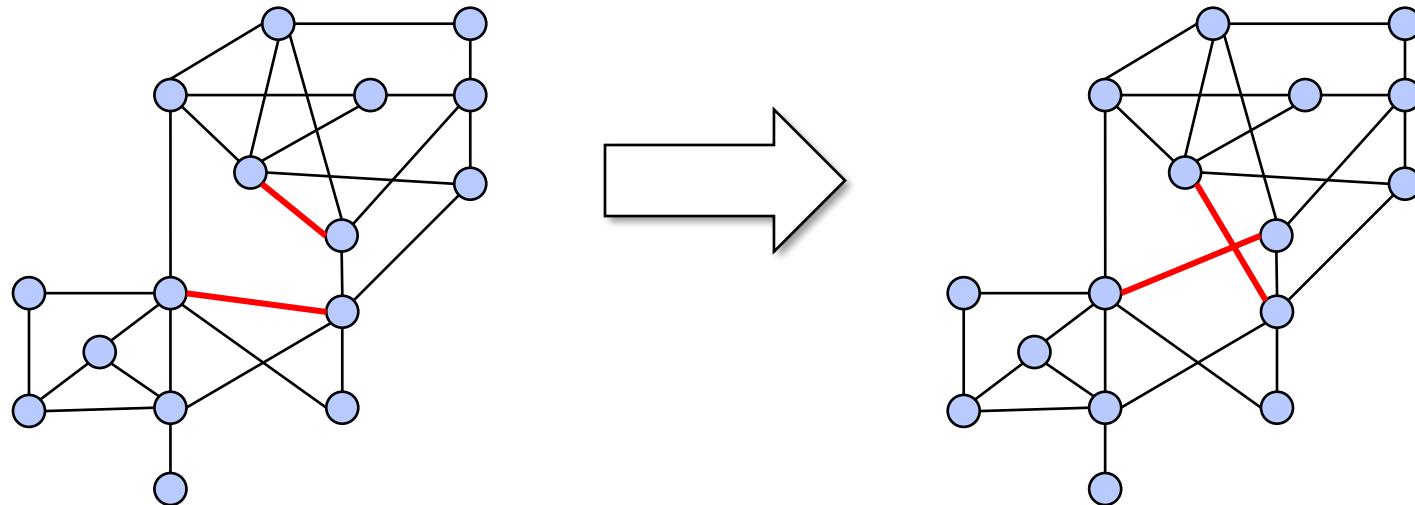
Modelo de configuración

- ◆ Se especifica la secuencia exacta del grado de cada nodo
- ◆ Se genera aleatoriamente un grafo que cumpla esta secuencia
- ◆ Algoritmo de generación muy sencillo
 1. Crear $g(u)$ “cabos sueltos” para cada nodo u
 2. Seleccionar dos cabos al azar y crear un arco
 3. Repetir 2 hasta unir todos los cabos

Modelo de configuración (cont)

- ◆ El procedimiento no es perfecto: se generan auto-enlaces y multi-enlaces
 - Según crece el tamaño del grafo el ratio de estos arcos es $\sim \text{avg}_u g(u)^2/n$
 - Los auto-enlaces y multi-enlaces se pueden omitir (o no) sobre la marcha en la generación
 - O se pueden rechazar cuando aparecen
 - O se puede hacer una fase final de intercambios de arcos para reducirlos lo más posible
- ◆ Otra opción, si se parte ya de un grafo no aleatorio con la densidad deseada
 - Recablear arcos aleatoriamente mediante swapping de pares de arcos, rechazando autoenlaces y enlaces múltiples

Modelo de configuración por recableado



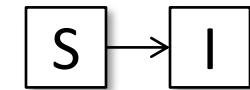
5. Procesos de difusión

Procesos de difusión en redes sociales

- ◆ Información (rumores, etc.), innovación, opiniones, modas, epidemias, etc.
- ◆ Patrones comunes para diferentes tipos de estado
- ◆ Las redes son proclives a efectos exponenciales (viralidad, cascadas)
- ◆ Los procesos de propagación son complejos –comprenderlos es un problema abierto y muy relevante
 - Marketing, control de epidemias, fake news, etc.

Difusión en redes sociales

- ◆ Modelo Susceptible-Infectado (SI)

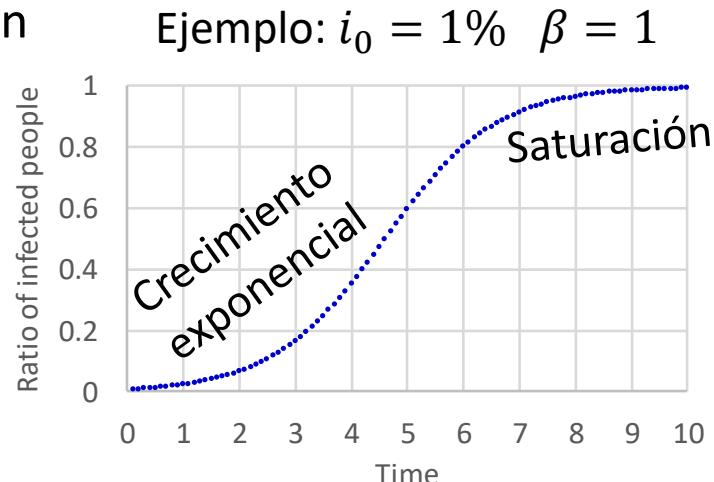


- Simplificación: grafo completo
- Individuos infectados y sanos (susceptibles)
- La probabilidad de encuentro entre dos personas cualesquiera es constante: β encuentros por persona y unidad de tiempo
- La probabilidad de contagio cuando una persona infectada se encuentra con una sana es 1 (equivalentemente, multiplicamos esta probabilidad en β)

- ◆ Dinámica

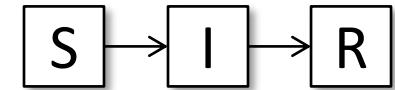
- En un instante t , i_t fracción de población infectada, $1 - i_t$ población sana
- Incremento esperado en la tasa de infección por unidad de tiempo \sim nº promedio de encuentros entre infectados y sanos

$$\frac{\Delta i_t}{\Delta t} = \beta i_t (1 - i_t)$$
$$\Rightarrow i_t = \frac{i_0 e^{\beta t}}{1 - i_0 + i_0 e^{\beta t}}$$



Difusión en redes sociales

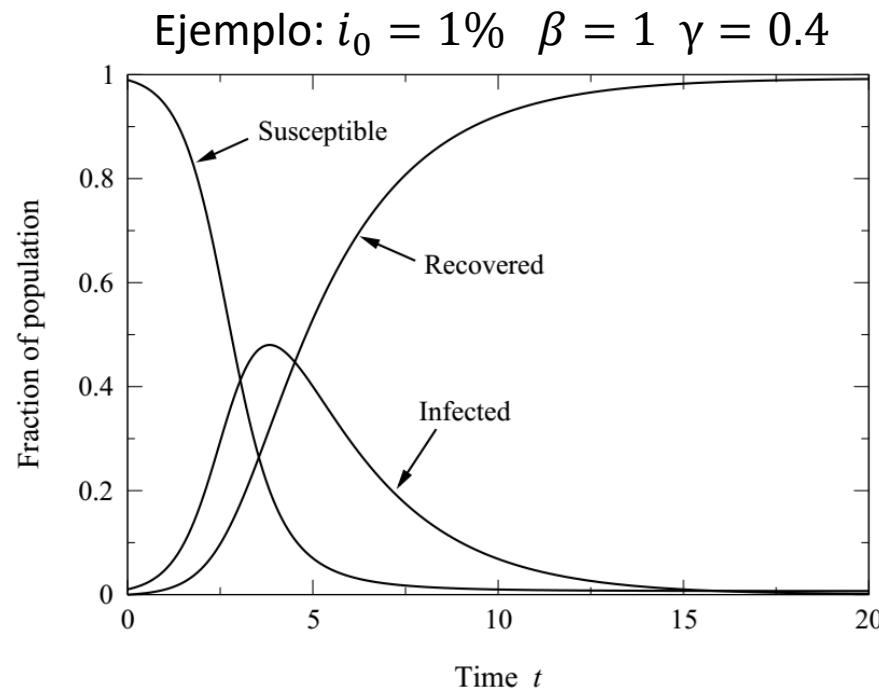
- ◆ **Modelo Susceptible-Infectado-Recuperado (SIR)**



- Igual que el modelo SI
- Salvo que en cada unidad de tiempo, γ personas infectadas se curan y quedan inmunes, o fallecen
- SI es un caso particular de SIR con $\gamma = 0$

- ◆ **Dinámica**

- Ecuaciones complejas, sólo pueden resolverse numéricamente
- Comportamiento asintótico $t \rightarrow \infty$
 - Si $\beta \leq \gamma$, la infección remite más rápido de lo que crece
 - Si $\beta > \gamma$, la infección primero crece y después remite
- $\beta = \gamma$ es la *transición epidémica*



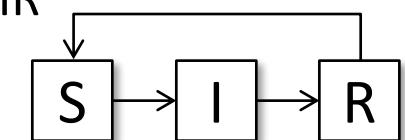
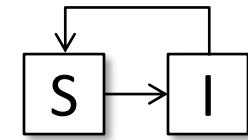
Difusión en redes sociales

- ◆ Modelo Susceptible-Infectado-Susceptible (SIS)
 - Igual que el modelo SIR
 - Salvo que los individuos recuperados se vuelven susceptibles de nuevo
 - Si es también un caso particular de SIS con $\gamma = 0$
- ◆ Dinámica

$$\frac{\Delta i_t}{\Delta t} = \beta i_t(1 - i_t) - \gamma i_t \quad \Rightarrow \quad i_t = \frac{(\beta - \gamma)i_0 e^{(\beta - \gamma)t}}{\beta(1 - i_0) - \gamma + \beta i_0 e^{(\beta - \gamma)t}}$$

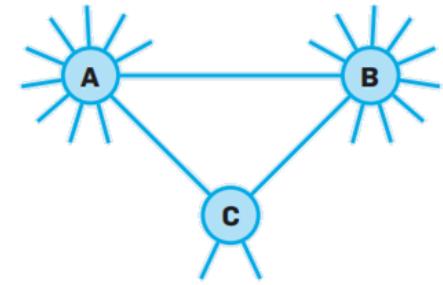
- Si $\beta > \gamma$ crecimiento logístico similar a SI, salvo $i_t \not\rightarrow 1$ con $t \rightarrow \infty$
En cambio, $i_t \rightarrow 1 - \gamma/\beta$, *estado de enfermedad endémica*: se infecta y recupera un número parecido de personas en cada momento
- Si $\beta < \gamma$ entonces $i_t \rightarrow 0$ exponencialmente como en SIR

- ◆ Más modelos: SIRS (tiempo δ de inmunidad), etc.



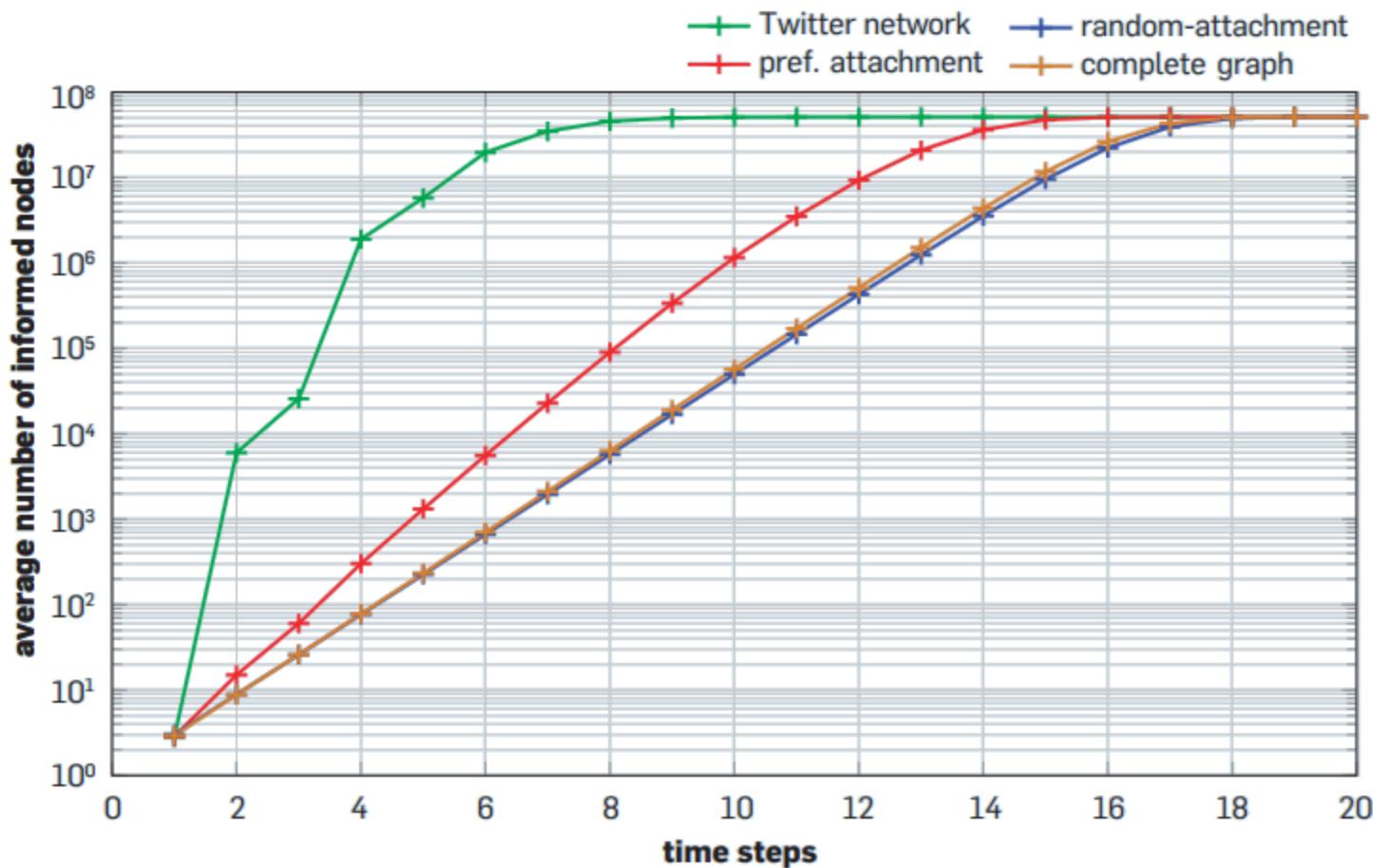
Difusión en redes sociales

- ◆ Para redes no completas, matemática muy compleja
 - Los modelos se analizan comúnmente por simulación
 - ◆ La dinámica de comunicación marca enormes diferencias
 - Cada cual con todos sus amigos
 - Cada cual con un amigo aleatorio
 - Unidireccional vs. bidireccional
 - Diferentes grados de susceptibilidad a la adopción
 - Propagación por múltiple exposición (presión social)
 - ...
 - ◆ Los nodos de bajo grado aceleran la propagación en algunos modelos
 - ◆ En redes no conexas el estado inicial puede marcar gran diferencia
 - P.e. según la infección se inicie o no en la componente gigante



“C pulls the rumor from A and quickly pushes it to B”

Difusión en redes sociales



B. Doerr, M. Fouz, T. Friedrich. Why rumors spread so quickly in social networks. Communications of the ACM 55(6), 2012

Software SNA

- ◆ Entorno interactivo: visualización, layout, métricas, plugins...
 - Gephi (también API Java)
- ◆ Manipulación de grafos, métricas, grafos aleatorios, visualización...
 - iGraph (C, extensiones R, Python, Ruby)
 - Jung (Java)
 - SNAP (C++)
- ◆ Modelado de dinámicas de red, animaciones
 - NetLogo (lenguaje propio)
 - SONIA (Java)
- ◆ ...

Algunos conjuntos de datos públicos

- ◆ Jure Leskovec @ Stanford University (SNAP)
 - <http://snap.stanford.edu/data>
- ◆ Paolo Boldi @ Università degli studi di Milano
 - <http://law.di.unimi.it/datasets.php>
- ◆ Gephi
 - <http://wiki.gephi.org/index.php/Datasets>
- ◆ Arizona State University
 - <http://socialcomputing.asu.edu/pages/datasets>
- ◆ Datamob
 - <http://datamob.org/datasets/tag/social-networks>
- ◆ ...