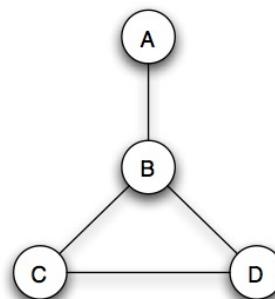


Grafi e rappresentazione delle Reti

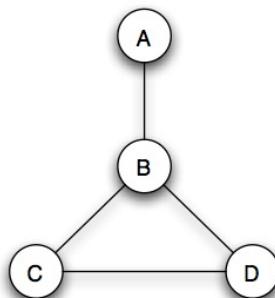
Capitolo 2

http://snap.stanford.edu/class/cs224w-readings/kleinber00book_ch02.pdf



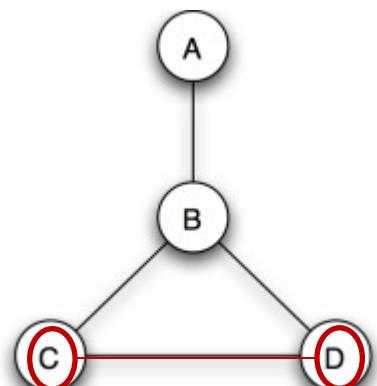
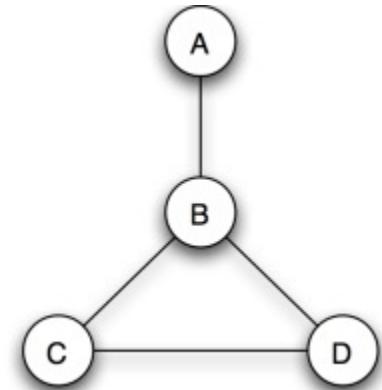
Grafi e rappresentazione delle Reti

Una buona scelta della rappresentazione della rete determina la nostra capacità di utilizzare il sistema con successo



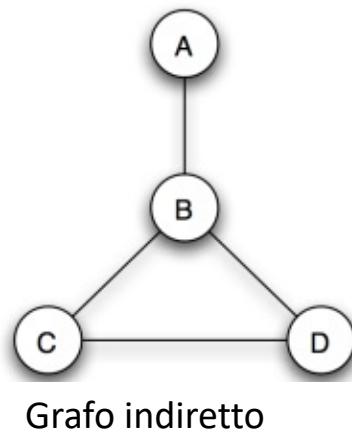
Definizione di Grafo

- Un grafo $G=(V,E)$ consiste di
 - Un insieme V di nodi (vertici)
 - Un insieme E di edge
 - ognuno collega due nodi
- Due nodi sono *vicini* (neighbor) se sono collegati da un edge
 - I nodi C e D sono adiacenti all'edge (C, D)

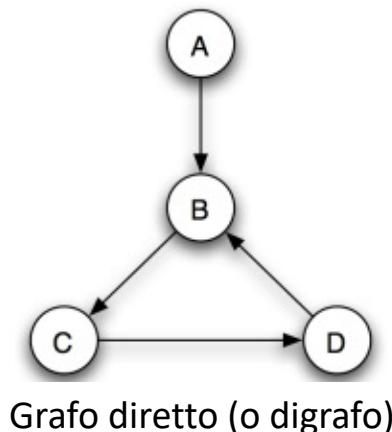


Grafi Diretti e Indiretti

- Una rete può essere orientata o non orientata.
 - Nelle reti non orientate, tutti i collegamenti (o edge) sono bidirezionali e l'ordine dei due nodi in un collegamento non ha importanza.
 - Nelle reti orientate, i collegamenti sono chiamati collegamenti orientati (o archi) e l'ordine dei nodi in un collegamento riflette la direzione: l'arco (i, j) va dal nodo di origine i al nodo di destinazione j .



Grafo indiretto

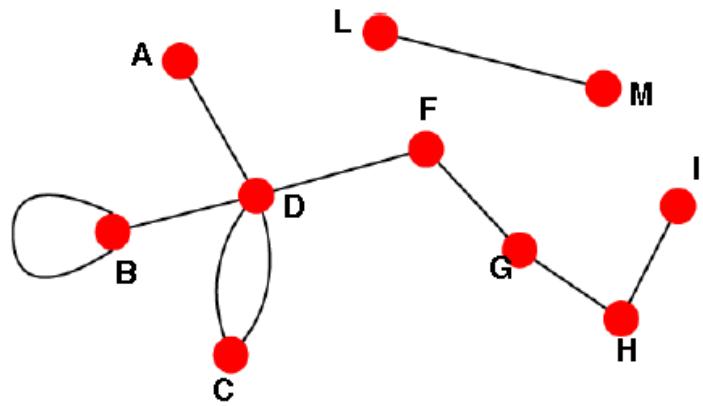


Grafo diretto (o digrafo)

- La differenza è sostanziale
 - Modelli differenti di formazione e mantenimento della rete
 - Algoritmi differenti

Grafi e Networks

Grafi diretti e indiretti codificano diversi tipi di reti



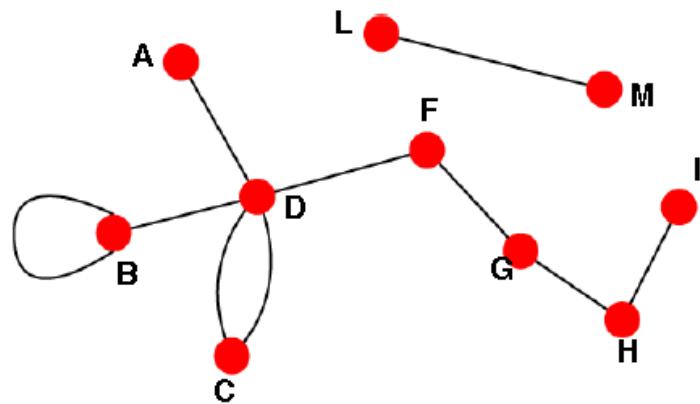
Grafi indiretti

La relazione tra due nodi è
frutto di un'azione concordata
tra i due nodi

**Es: amicizia, alleanza,
conoscenza, connessione, ecc.**

Grafi e Networks

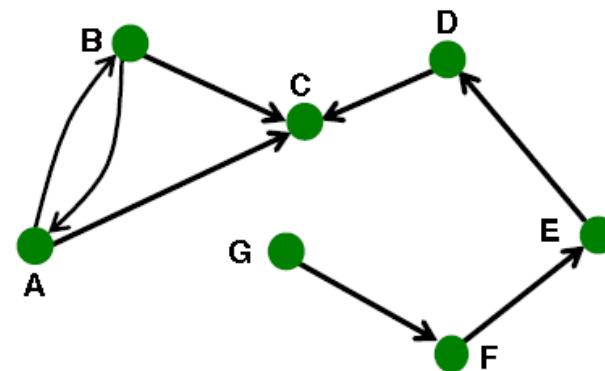
Grafi diretti e indiretti codificano diversi tipi di reti



Grafi indiretti

La relazione tra due nodi è frutto di un'azione concordata tra i due nodi

Es: amicizia, alleanza, conoscenza, connessione, ecc.



Grafi diretti

La relazione tra due nodi è frutto di un atto unilaterale

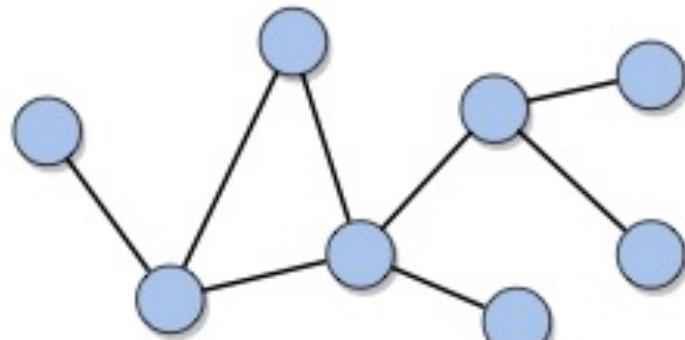
Es: link a pagine web, followers, citazioni di un articolo

Reti Pesate

- Una rete può essere non pesata o non.
 - In una rete pesata, i collegamenti hanno pesi associati: il collegamento pesato (i, j, w) tra i nodi i e j ha peso w .

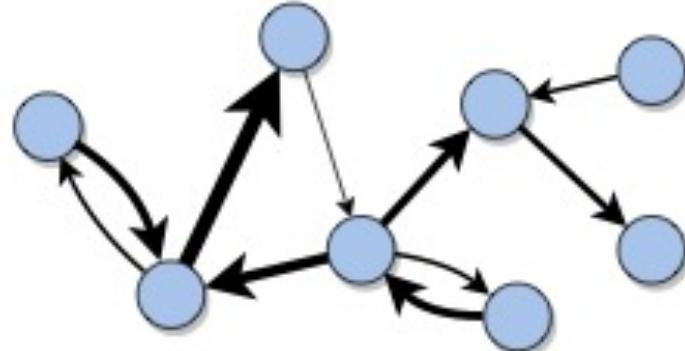
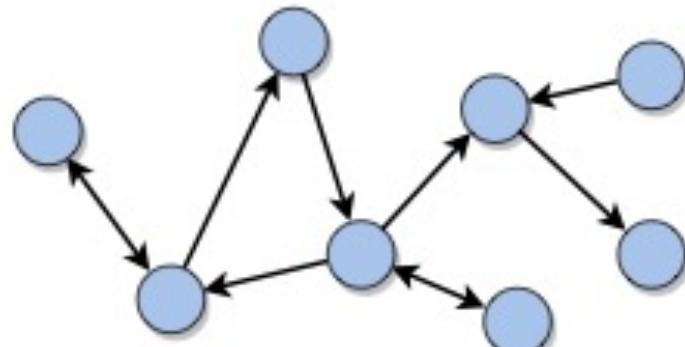
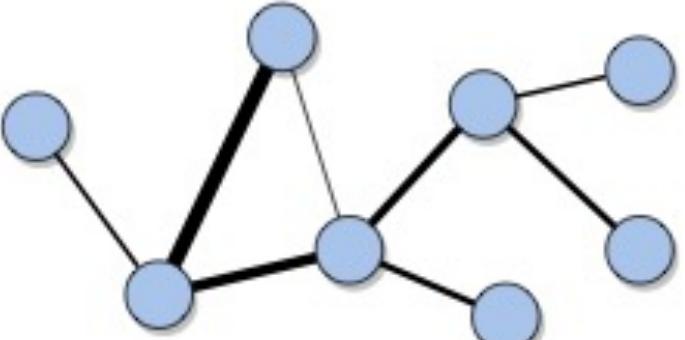
Undirected

Unweighted

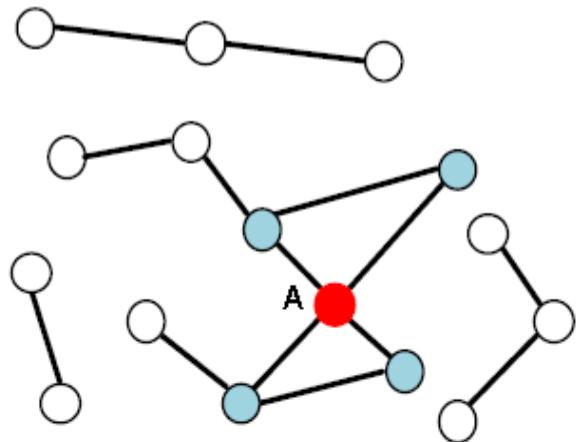


Directed

Weighted



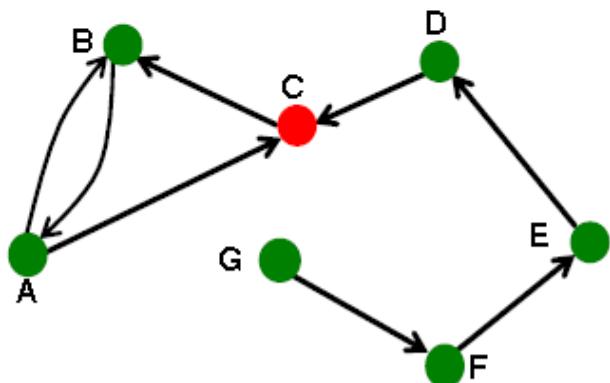
Grado di un nodo



Grado del nodo i: numero k_i di archi adiacenti ad i

Grado medio:

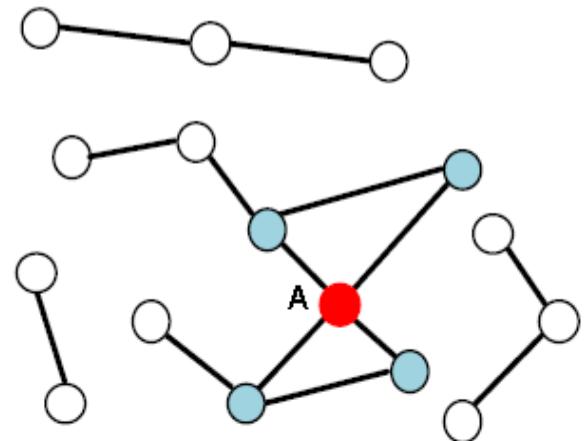
$$\bar{k} = \langle k \rangle = \frac{1}{N} \sum_{i=1}^N k_i = \frac{2E}{N}$$



In grafi orientati si distingue ulteriormente tra
in-degree: numero di archi entranti
out-degree numero di archi uscenti

Intorno (Neighborhood)

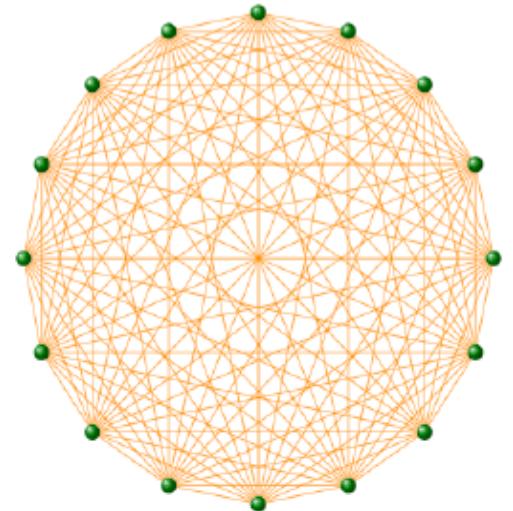
- L'*intorno* di un nodo è l'insieme di tutti i suoi vicini
- L'intorno di un insieme di vertici S è l'insieme dei vertici del grafo che non appartengono ad S ma sono adiacenti a vertici di S



Grafo completo

Il massimo numero di edge in un grafo non orientato di N nodi è

$$E_{\max} = \binom{N}{2} = \frac{N(N - 1)}{2}$$



Un grafo non orientato di n nodi con tale numero di edge si dice **completo**

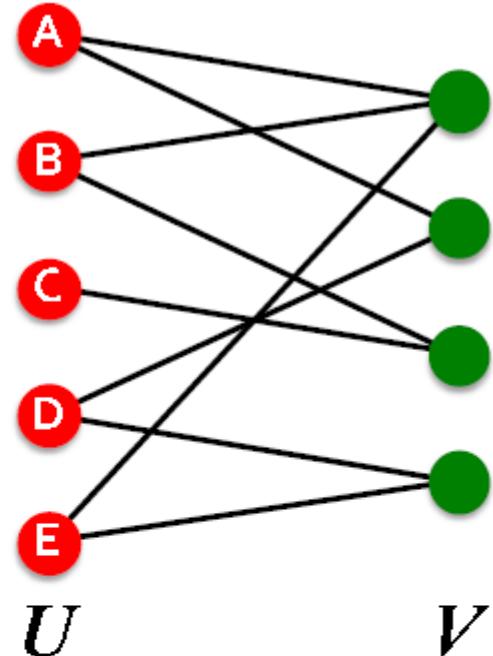
Il grado di ogni nodo (e quindi il grado medio) è N-1

Grafo bipartito

Un grafo si dice **bipartito** se i nodi possono essere partizionati in due insiemi disgiunti U e V in modo tale che esistono solo edge che connettono un nodo in U ed uno in V , cioè U e V sono **insiemi indipendenti**

Es.

- Autore—articolo(di cui è autore)
- Film—attore(che ha partecipato al film)

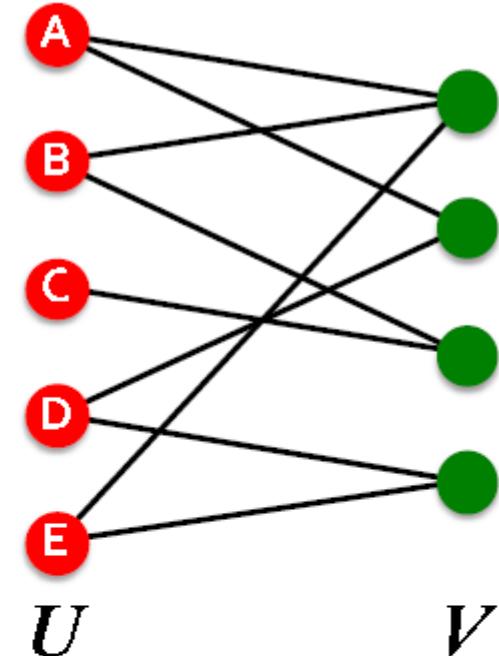


Grafo bipartito

Un grafo si dice **bipartito** se i nodi possono essere partizionati in due insiemi disgiunti U e V in modo tale che esistono solo edge che connettono un nodo in U ed uno in V , cioè U e V sono **insiemi indipendenti**

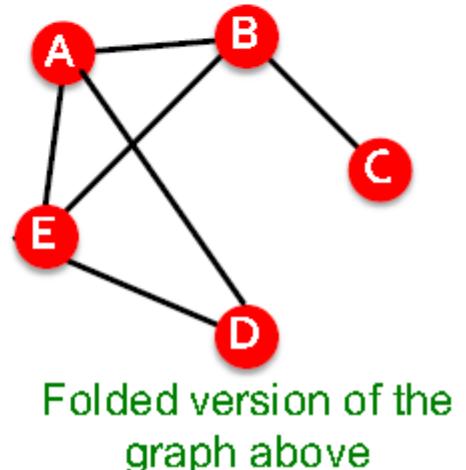
Es.

- Autore—articolo(di cui è autore)
- Film—attore(che ha partecipato al film)



Grafo Folded:

- Rete delle collaborazioni tra autori
- Copresenze di attori in film

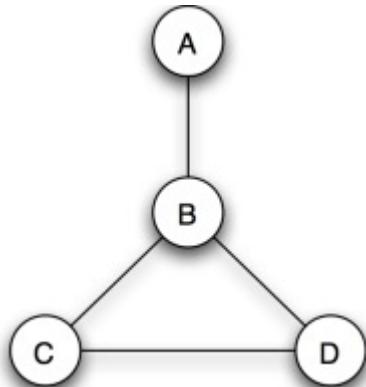


Folded version of the graph above

Rappresentazione dei Grafi

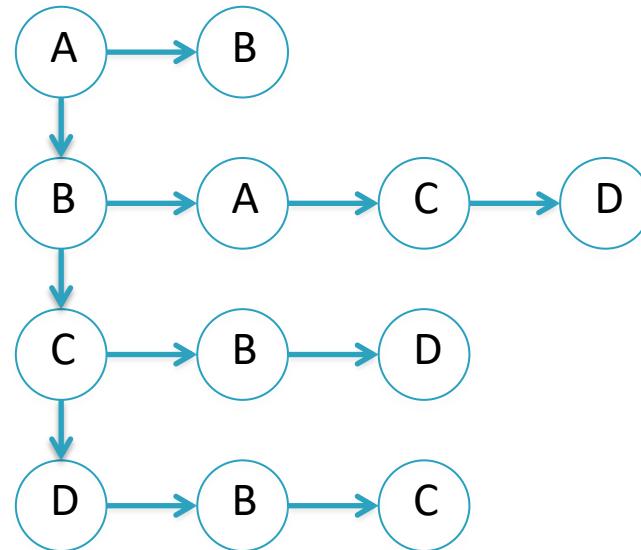
- Un grafo è una coppia di insiemi
 - $G = (V, E)$
 - V = insieme dei vertici (nodi)
 - E = insieme degli edge/archi
- Rappresentazioni più utilizzate
 - Matrice di adiacenza
 - Matrice $n \times n$ (n numero dei vertici in V)
 - Elemento (i, j) vale 1 se esiste l'edge tra i e j
 - Vale $w_{i,j}$ se il grafo è pesato e w è la funzione peso
 - Lista dei vertici V e degli edge E
 - Per ogni vertice v abbiamo la lista dei vertici adiacenti a v
 - Rappresentazione grafica

Rappresentazione di un Grafo: Liste delle adiacenze



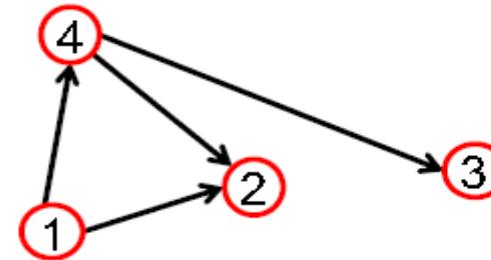
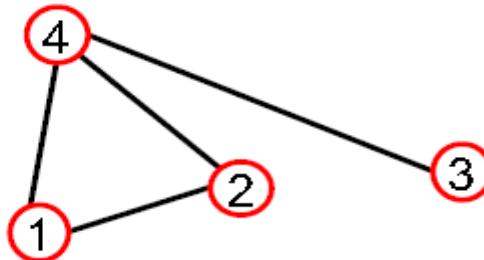
	A	B	C	D
A	0	1	0	0
B	1	0	1	1
C	0	1	0	1
D	0	1	1	0

Matrice di
adiacenza



liste di
adiacenza

Rappresentazione di un Grafo: Matrice delle adiacenze



$A_{ij} = 1$ if there is a link from node i to node j

$A_{ij} = 0$ otherwise

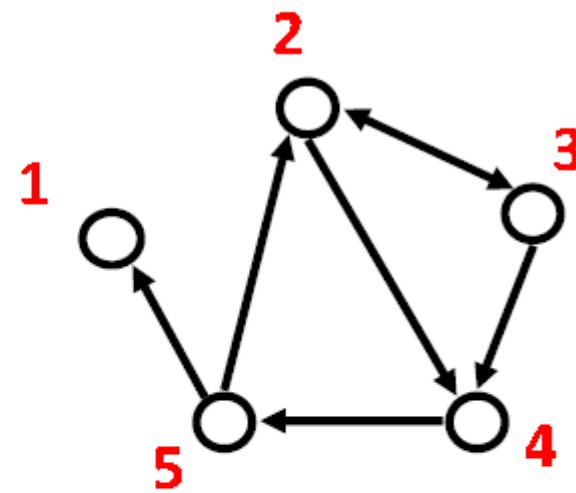
$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

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

Note that for a directed graph (right) the matrix is not symmetric.

Rappresentazione di un Grafo: Lista degli archi

- (2, 3)
- (2, 4)
- (3, 2)
- (3, 4)
- (4, 5)
- (5, 2)
- (5, 1)



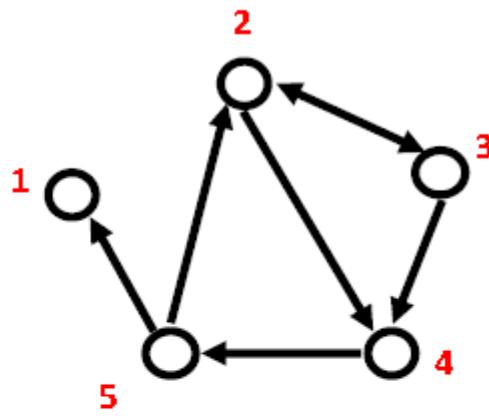
Liste delle ediacenze

Più semplici da utilizzare quando la rete è

- grande
- sparsa

Permette di trovare velocemente tutti i vicini di un nodo dato

- 1:
- 2: 3, 4
- 3: 2, 4
- 4: 5
- 5: 1, 2



Network reali sono grafi sparsi

$$E \ll E_{\max} \text{ (or } \bar{k} \ll N-1)$$

WWW (Stanford-Berkeley):	$N=319,717$	$\langle k \rangle = 9.65$
Social networks (LinkedIn):	$N=6,946,668$	$\langle k \rangle = 8.87$
Communication (MSN IM):	$N=242,720,596$	$\langle k \rangle = 11.1$
Coauthorships (DBLP):	$N=317,080$	$\langle k \rangle = 6.62$
Internet (AS-Skitter):	$N=1,719,037$	$\langle k \rangle = 14.91$
Roads (California):	$N=1,957,027$	$\langle k \rangle = 2.82$
Proteins (S. Cerevisiae):	$N=1,870$	$\langle k \rangle = 2.39$

(Source: Leskovec et al., Internet Mathematics, 2009)

La matrice delle adiacenze contiene moltissimi zeri!

- densità (E/E_{\max}): WWW= 1.51×10^{-5} , MSN IM = 2.27×10^{-8})

Grafi Pesati o Segnati

- Possiamo associare agli archi delle informazioni aggiuntive
 - Peso (lunghezza del collegamento)
 - Tipo (amico, parente, collega,)
 - forza del legame/ranking (miglior amico, secondo miglior amico,...)
 - ritardo (tempo di trasmissione)
 - affidabilità (percentuale di errore nella trasmissione)
 - costo (costo di utilizzo del collegamento)
- In un grafo *pesato* ogni arco ha un numero associato che ne definisce il peso
- In un grafo *segnato* ogni arco ha un segno positivo o negativo (es. amici-nemici)

Vertici ed Archi

- Nello studio delle reti i nodi e gli archi rappresentano entità del mondo reale
 - alcune astrazioni di network comunemente utilizzate

WWW >> directed multigraph with self-edges

Facebook friendships >> undirected, unweighted

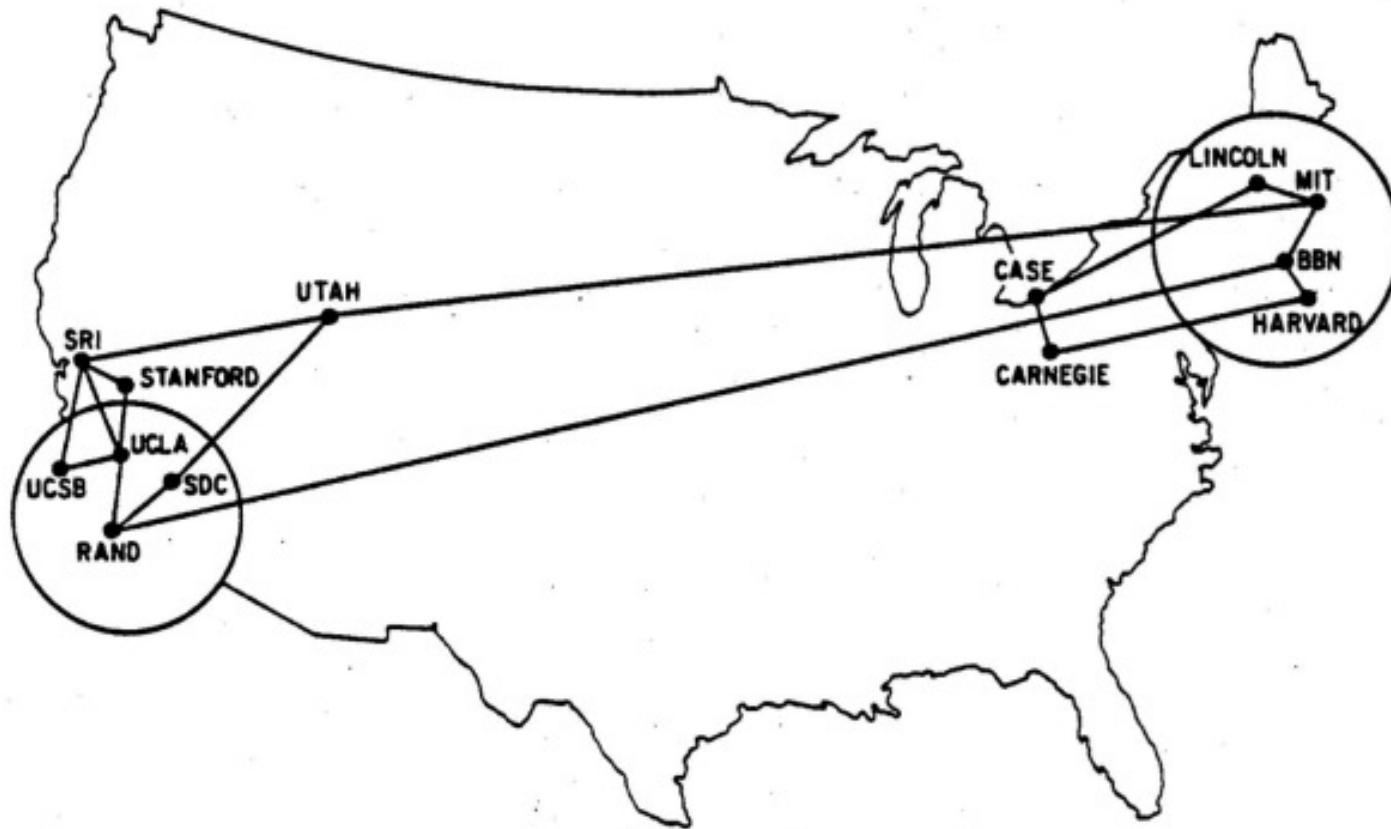
Citation networks >> unweighted, directed, acyclic

Collaboration networks >> undirected multigraph or weighted graph

Mobile phone calls >> directed, (weighted?) multigraph

Protein Interactions >> undirected, unweighted with self-interactions

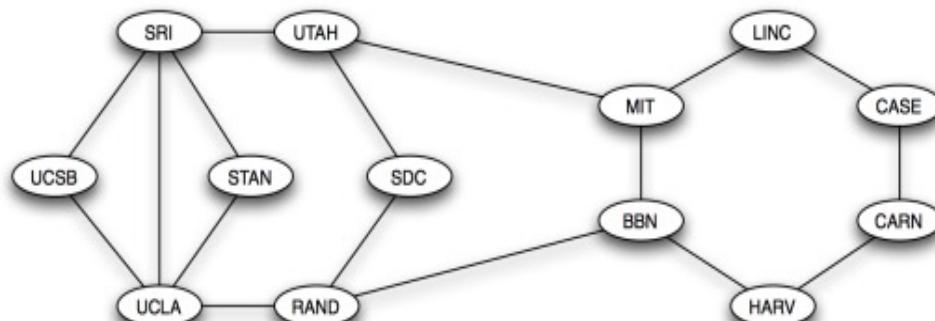
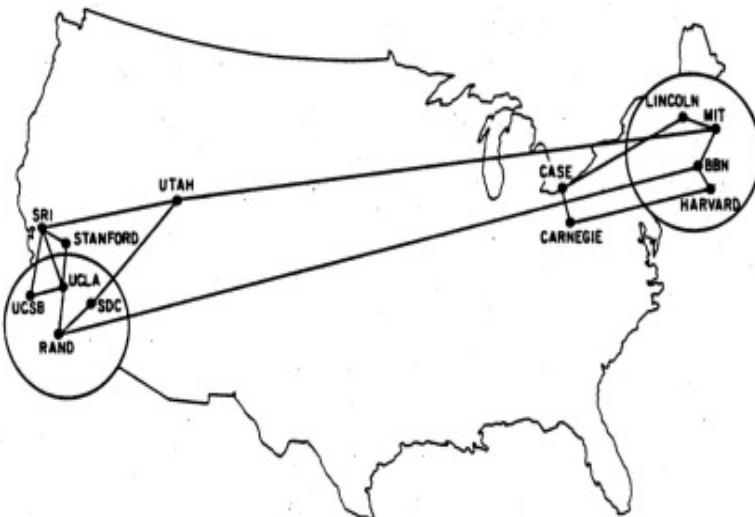
ARPANET: Precursore di Internet



- Creata nel 1970 con 13 nodi

Il Grafo di ARPANET

- Siamo interessati solo alla connettività
 - Le distanze possono essere rappresentate come pesi



Reti di Trasporto



- La terminologia dei grafi è in gran parte derivata dal mondo dei trasporti
 - “cammino minimo”, “diametro”, “flusso”

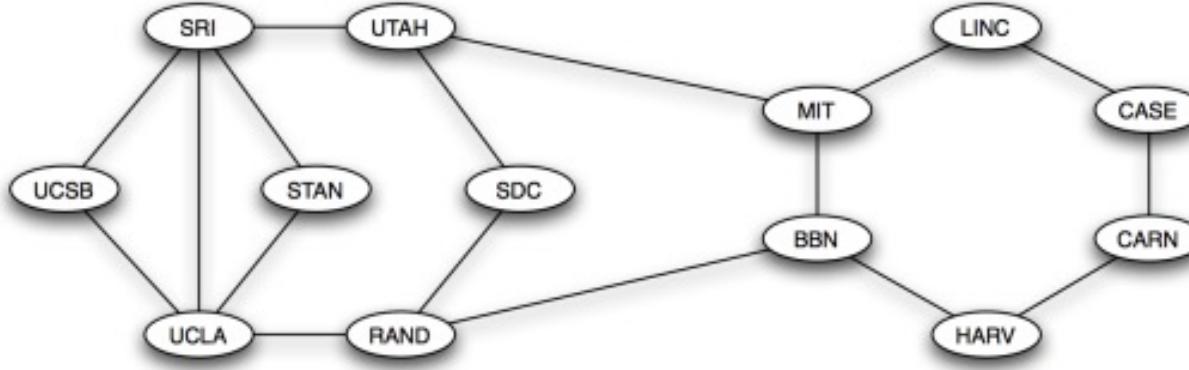


Cammini

- Un elemento caratteristico delle reti è che i nodi possono influenzarsi tramite relazioni indirette
- Diverse cose spesso viaggiano lungo gli archi del grafo
 - Mezzi di trasporto
 - informazioni
 - influenza
 - malattie
- Un **cammino (path)** è una sequenza di nodi con la proprietà che ogni coppia di nodi consecutivi del cammino sono connessi da un edge
 - Se tra due nodi esiste un cammino allora i due nodi sono in relazione indiretta tra loro



Cammini

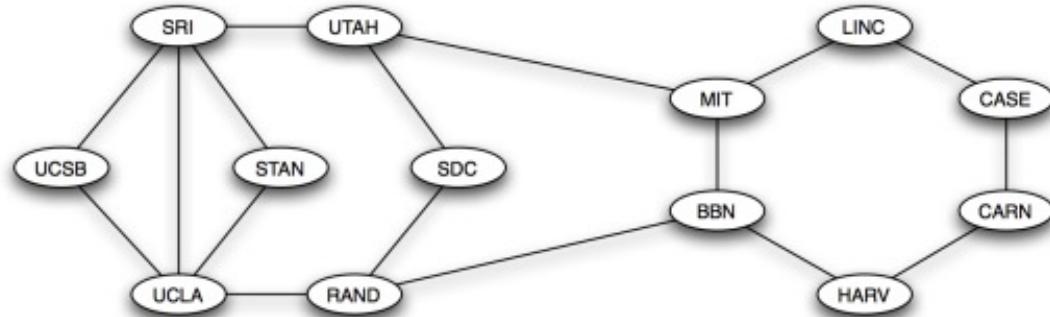


- MIT – BBN – RAND – UCLA è un cammino
- UCSB – UCLA – RAND – MIT non è un cammino
- Un percorso può attraversare lo stesso nodo più volte
 - SRI – STAN – UCLA – SRI – UTAH – MIT
- Un **cammino semplice** non attraversa mai uno stesso nodo due volte
- Un **cammino minimo** è un cammino che attraversa il minor numero possibile di archi

Cicli

Ciclo: cammino semplice che inizia e termina nello stesso nodo

- LINC – CASE – CARN – HARV – BBN – MIT – LINC è un ciclo
- Ogni ciclo ha almeno
 - tre archi



- Nelle *reti di comunicazione* o trasporto ogni nodo appartiene ad uno o più cicli
 - Ridondanza inserita per aumentare la robustezza della rete e garantire il funzionamento della rete anche in presenza di guasti
- In una *rete sociale* i cicli sono frequenti ma non voluti
 - Es: l'amica della cugina di mia moglie è sorella del mio collega d'ufficio

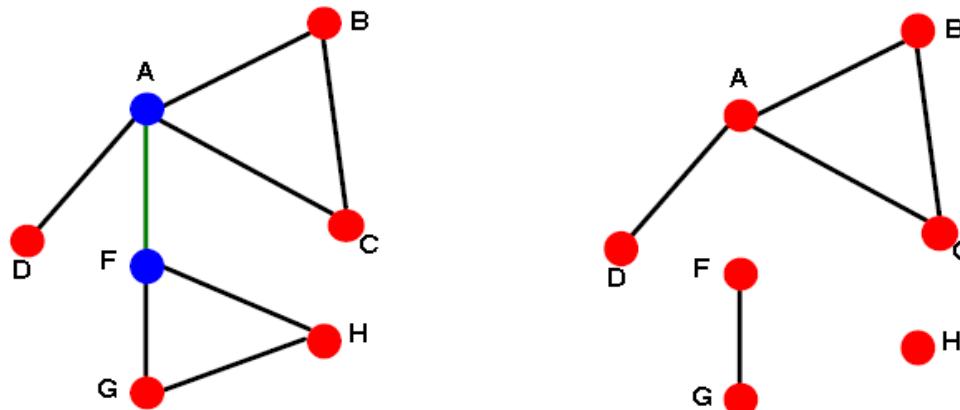
Cammini e Cicli Diretti

- Cammini e cicli diretti si definiscono in maniera analoga al caso indiretto
 - Bisogna tener conto delle direzioni degli archi del cammino
- A volte possiamo considerare cicli indiretti anche in grafi diretti
 - Ignoriamo la direzione dell'arco
 - Utile se ci interessa l'esistenza di una relazione, indipendentemente da chi l'ha attivata

Connettività

Grafo (non orientato) connesso: ogni coppia di vertici è unita da un cammino

Un grafo non connesso è formato da due o più componenti connesse



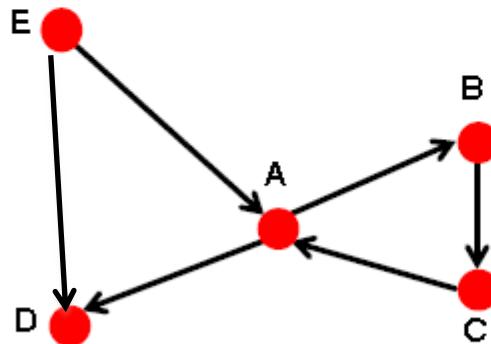
Bridge: edge che se rimosso disconnette il grafo

Punto di articolazione: nodo che se rimosso (insieme agli edge adiacenti) disconnette il grafo

Connettività

Grafo orientato **fortemente connesso**: ogni coppia ordinata di vertici è unita da un cammino
(es. vertici A e B: esistono sia path da A a B che da B ad A)

NON fortemente connesso:



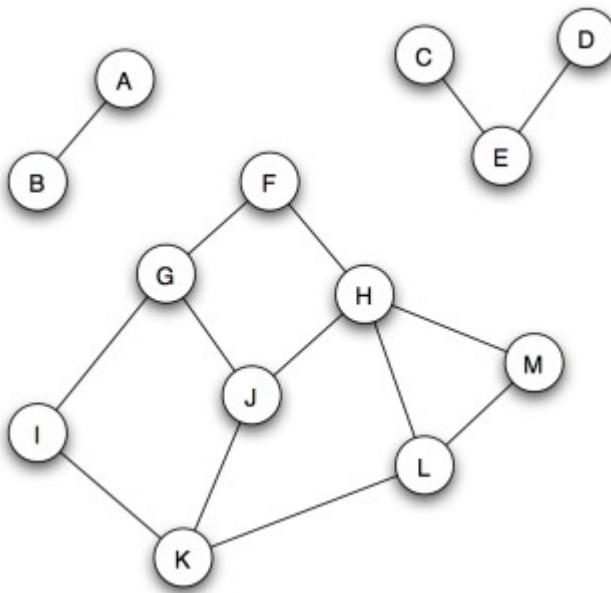
Componenti

- Se un grafo non è connesso può essere diviso in sottografi che sono a loro volta connessi
- Una **componente连通子图** di un grafo non diretto è un sottinsieme di nodi tali che
 - Ogni nodo ha un cammino verso ogni altro nodo della sua componente
 - Per ogni nodo u non appartenente alla componente esiste un nodo v nella componente tale che non esiste nessun cammino da u a v
- Un arco di una componente è un **bridge** se la sua cancellazione sconnette la componente

Componenti

- Se un grafo non è connesso può essere diviso in sottografi che sono a loro volta connessi
- Una ***componente connessa*** (risp. fortemente connessa) di un grafo non diretto (risp. grafo diretto) è un sottinsieme di nodi tali che
 - Ogni nodo ha un cammino (risp. cammino diretto) verso ogni altro nodo della sua componente
 - Per ogni nodo u non appartenente alla componente esiste un nodo v nella componente tale che non esiste nessun cammino da u a v

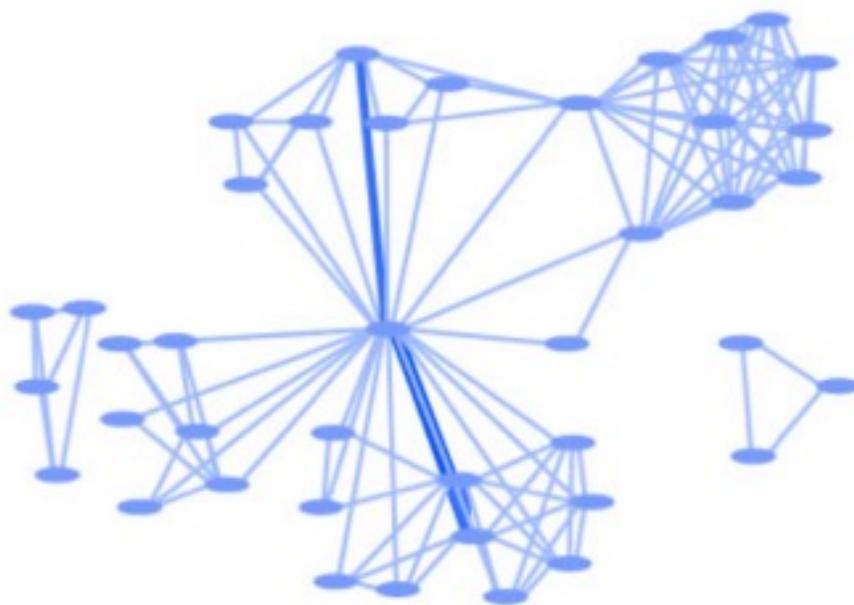
Componenti



- Tre componenti connesse
 - $\{A, B\}$, $\{C, D, E\}$, $\{F, G, \dots, M\}$
- $\{H, L, M\}$ non è una componente
- L'arco (D, E) è un bridge

Analisi delle Componenti

- In molte situazioni è importante riuscire ad individuare le componenti connesse di una rete

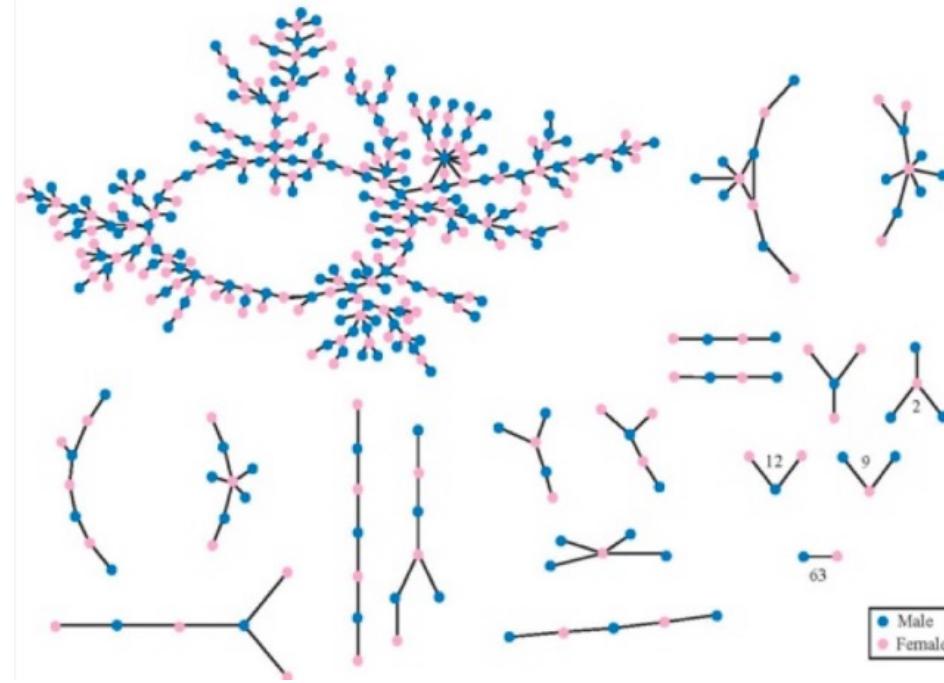


Grafo delle collaborazioni in un centro di ricerca

Componenti Giganti

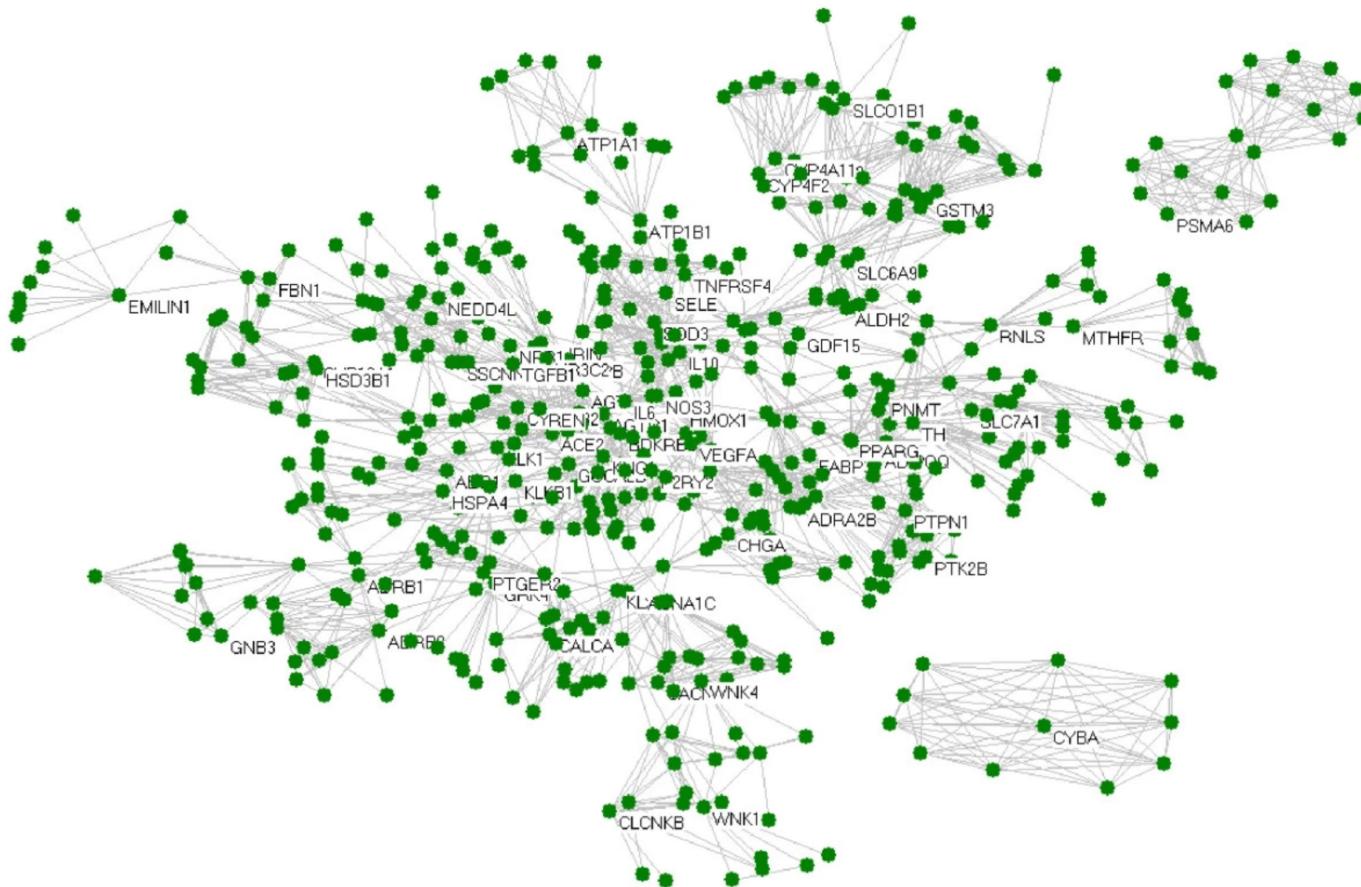
- Molte reti non sono connesse
ma contengono una *componente gigante*
 - una componente che contiene una percentuale
molto alta dei nodi della rete

Componenti Giganti



- Romantic relationships in an American high school over an 18-month period.
- Edges were not all present at once; rather, there is an edge between two people if they were romantically involved at any point during the time period.

From: [Construction and analysis of the protein-protein interaction network related to essential hypertension](#)

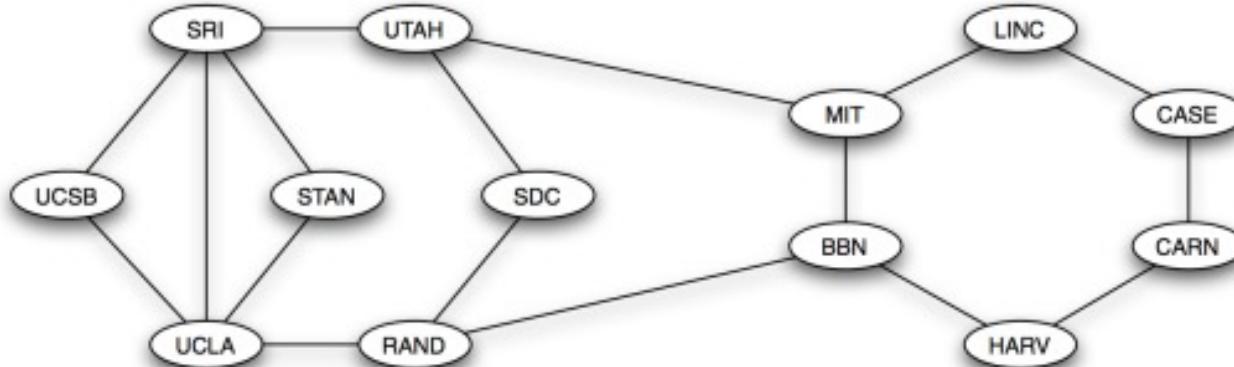


Ricapitolando:Caratteristiche principali delle reti reali

Componente gigante: esistenza componente connessa “*molto grande*”

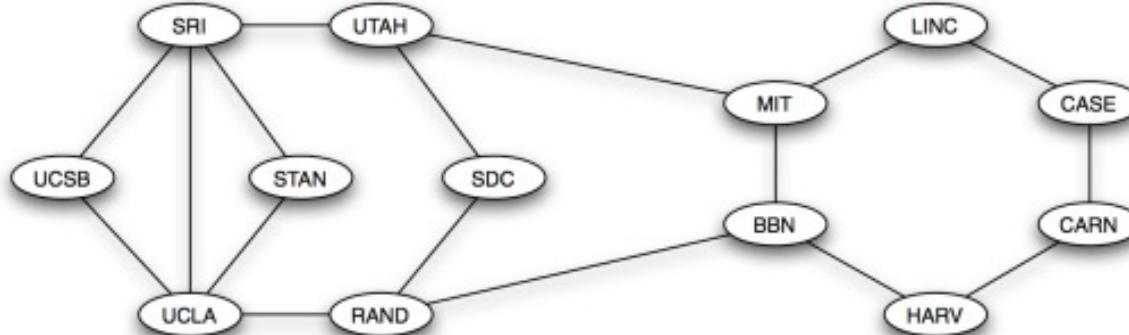
Distanze

- La *distanza* tra una coppia di nodi è la lunghezza del cammino minimo che li unisce
 - Utilizzando il concetto di lunghezza di un cammino, possiamo parlare nodi vicini o lontani in un grafo.
- Il *diametro* di un grafo è la più grande distanza tra coppie di nodi del grafo
 - Qual è la distanza tra MIT e SDC?
 - Qual è il diametro della rete?



Calcolo delle distanze minime

- Dato un grafo, come troviamo le distanze minime *da un nodo a tutti gli altri?*
 - Abbiamo bisogno di un metodo sistematico per determinare le distanze
- Come possiamo approcciare il problema?
 - Una possibilità è la ricerca breadth-first

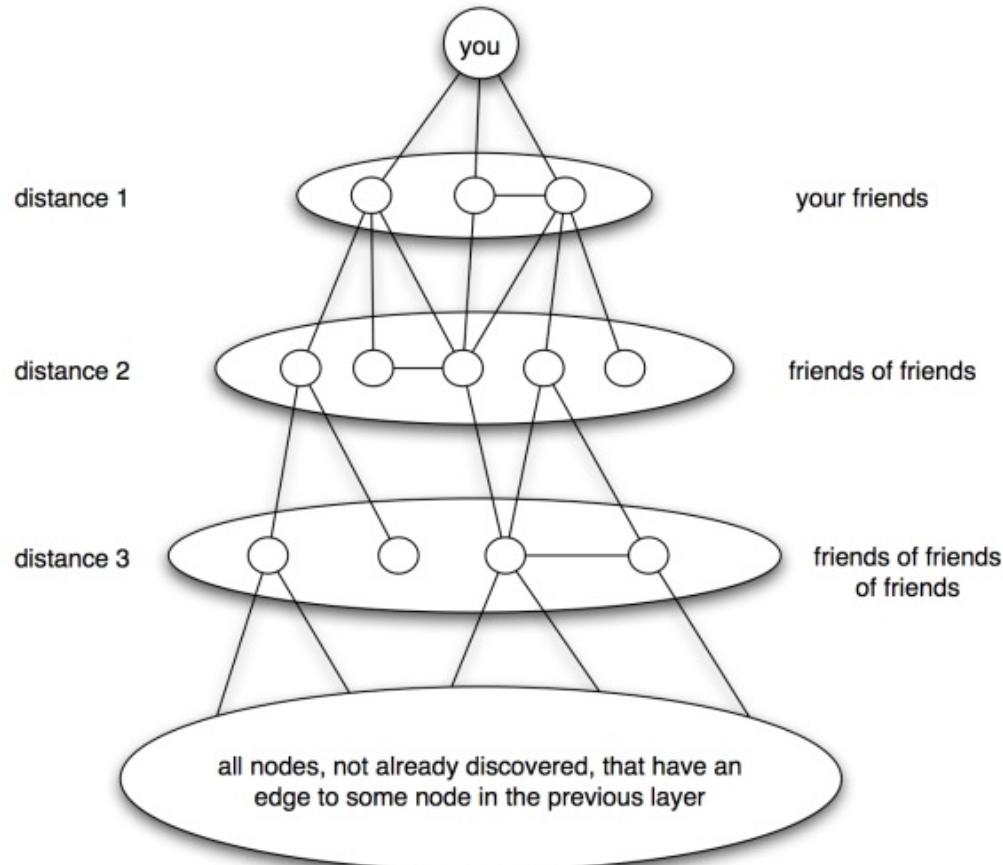


Breadth-first search (BFS)

- Dal nodo sorgente(*root*)
 - Trova tutti i nodi direttamente connessi
 - Questi sono i nodi a “distanza 1”
 - Trova tutti i nodi che sono direttamente connessi ai nodi a distanza 1 e che non sono ancora stati visitati
 - Questi sono i nodi a “distanza 2”
 - ...
 - Trova tutti i nodi che sono direttamente connessi a nodi a distanza j e che non sono già stati visitati
 - Questi nodi sono a distanza $j+1$ ”

BFS

- scopre distanze tra i nodi, un *layer (strato)* alla volta



- ogni layer è costituito dai nodi che hanno un almeno un vicino nello strato precedente.

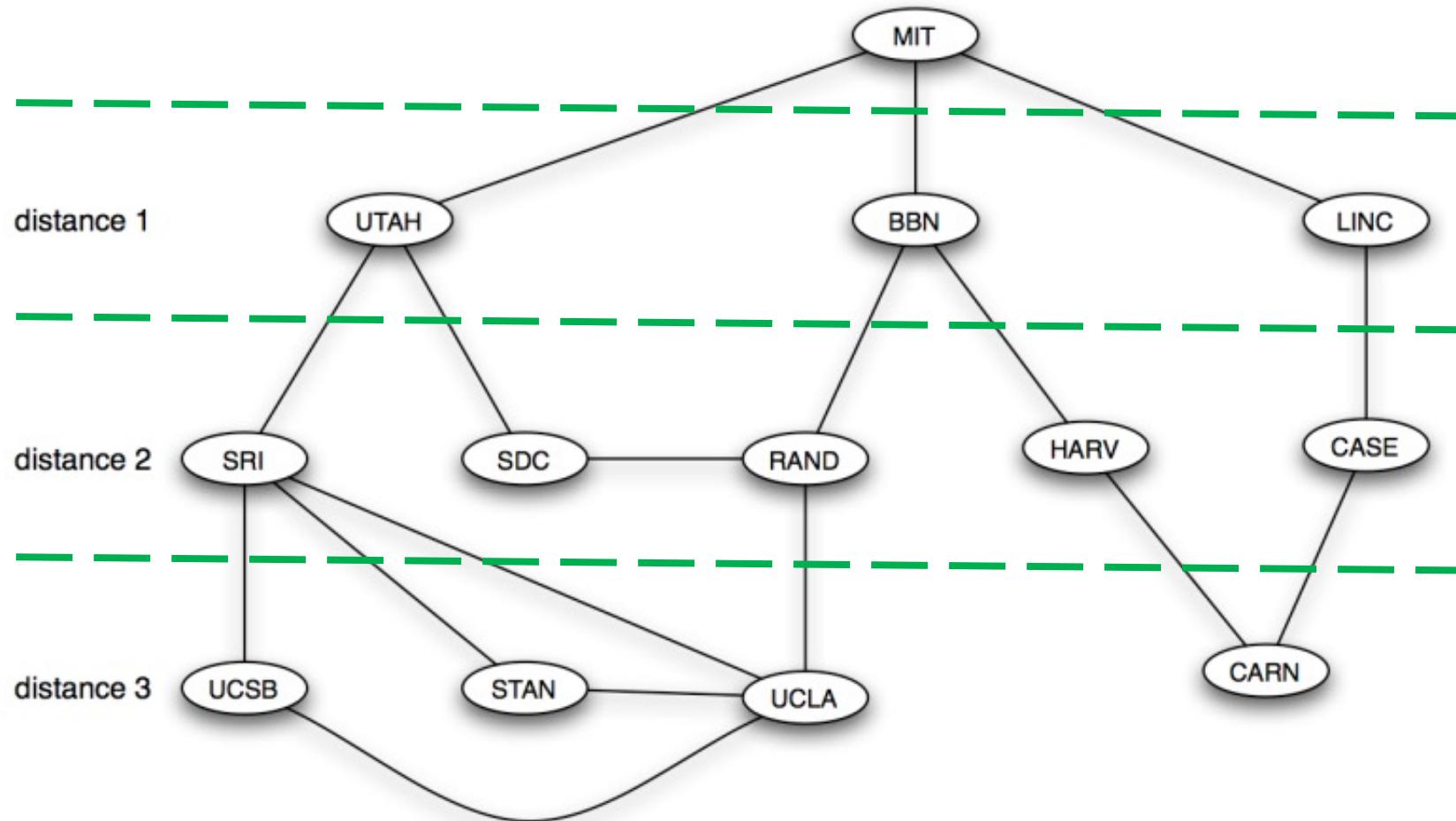
BFS: ALGORITMO

- 1) Dichiara tutti i tuoi amici essere a distanza 1.
- 2) Trova tutti i loro amici (senza contare quelli che sono già amici tuoi), e dichiara che questi sono a distanza di 2.
- 3) Trova tutti i loro amici (di nuovo, senza contare le persone già trovate ad una distanza ≤ 2) e dichiarale essere a distanza di 3.
- 4) Continua in questo modo; effettuando la ricerca in strati successivi, ciascuno rappresentante la distanza successiva.

Ogni nuovo livello è costruito da tutti quei nodi che

- non sono già stati scoperti in strati precedenti, e
- hanno un vicino nello strato precedente.

BFS sul Grafo di Arpanet (dicembre 1970) partendo da MIT



Small world phenomenon

In termini informali, si parla di effetto *Small world* quando si riscontrano distanze sorprendentemente *brevi* tra coppie di nodi

Small world phenomenon

In termini informali, si parla di effetto ***Small world*** quando si riscontrano distanze sorprendentemente *brevi* tra coppie di nodi

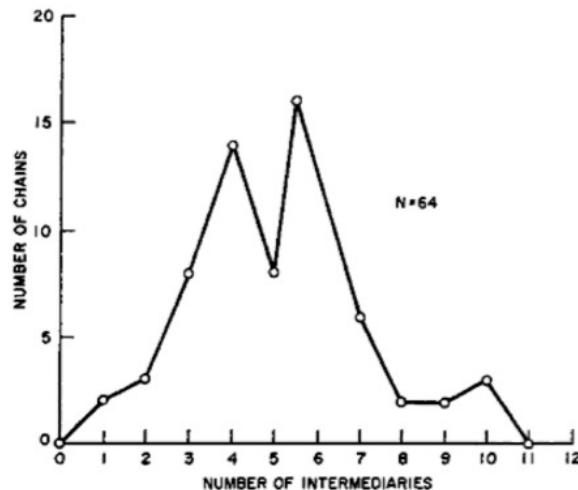
Primo esperimento condotto da Stanley Milgram negli anni 1960

- A 296 persone scelte a caso negli USA fu chiesto di far arrivare una lettera ad un certo destinatario
spedendola ad una persona che conoscevano direttamente
 - Ogni partecipante conosceva un profilo del destinatario (residenza, occupazione, tipo di studi, città di origine, ecc.)
 - il numero medio di passaggi intermedi delle lettere che arrivarono a destinazione risultò sorprendentemente *breve*

Small world phenomenon

Sei Gradi di Separazione

- Nell'esperimento di Milgram 64 lettere arrivarono a destinazione con una lunghezza media del percorso inferiore a 6



- Diversi esperimenti sono stati condotti che confermano i risultati di Milgram
- Grandi reti sociali hanno distanze “brevi”

Ricapitolando:Caratteristiche principali delle reti reali

Componente gigante: esistenza di una componente connessa “*molto grande*”

Fenomeno Small-world : distanze “brevi” tra coppie di nodi
(in una stessa componente connessa)

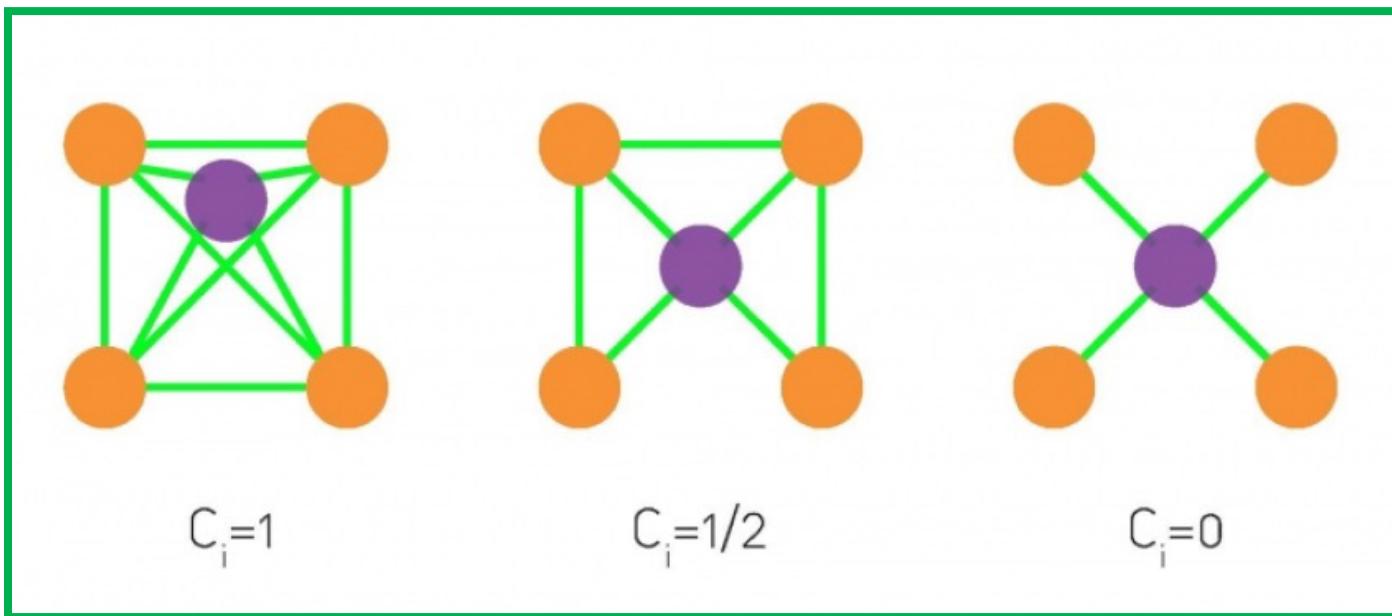
Clustering

Il **coefficiente di clustering** indica quanto i vicini di un nodo sono connessi tra loro.

Per un nodo i di grado k_i , il coefficiente di **clustering locale** è definito come

$$C_i = \frac{2L_i}{k_i(k_i-1)}$$

dove L_i è il numero di edge esistenti tra i vicini del nodo i .



Clustering

Il **coefficiente di clustering** indica quanto i vicini di un nodo sono connessi tra loro.

Per un nodo i di grado k_i , il coefficiente di **clustering locale** è definito come

$$C_i = \frac{2L_i}{k_i(k_i-1)}$$

dove L_i è il numero di edge esistenti tra i vicini del nodo i .

- C_i è sempre compreso tra 0 e 1:
 - $C_i = 0$ se tutti i vicini di i sono indipendenti tra loro.
 - $C_i = 1$ se i vicini del nodo i formano un grafo completo
- C_i rappresenta la probabilità che due vicini di i scelti a caso siano connessi da edge
- In sintesi, C_i misura la densità collegamento locale della rete: più il vicinato di i è densamente interconnesso, maggiore è il suo coefficiente di clustering locale

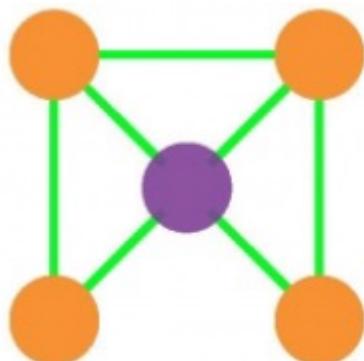
Clustering

Il **coefficiente di clustering** di un'intera rete viene catturato dal coefficiente di clustering medio

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^N C_i$$

che rappresenta la media dei coefficienti di clustering locali C_i , sui nodi $i = 1, \dots, N$,

$\langle C \rangle$ rappresenta la probabilità che selezionando casualmente due vicini di un qualche nodo, questi hanno un edge che li connette.



$$(1/2 + 2/3 + 2/3 + 1 + 1)/5 = 19/30$$

Ricapitolando:Caratteristiche principali delle reti reali

Componente gigante: esistenza di una componente connessa “*molto grande*”

Fenomeno Small-world : distanze “brevi” tra coppie di nodi
(in una stessa componente connessa)

Coefficiente Clustering $\langle C \rangle = \frac{1}{N} \sum_{i=1}^N C_i$ “*elevato*”

$\langle C \rangle$ rappresenta la probabilità che selezionando casualmente due vicini di un qualsiasi nodo, questi hanno un edge che li connette.

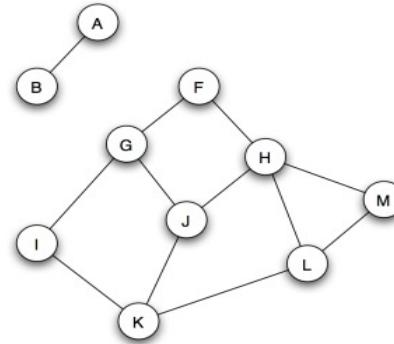
Distribuzione dei gradi

Ricorda: il grado di un nodo corrisponde al numero di vicini di quel nodo

Consideriamo una rete e
sia p_d la frazione di nodi che hanno grado d .

Ad esempio, nella rete a destra abbiamo:

$$p_1=2/10 \quad p_2=3/10 \quad p_3=4/10 \quad p_4=1/10$$



Questo rapporto rappresenta la probabilità che un nodo scelto in maniera casuale abbia quel grado

Distribuzione dei gradi

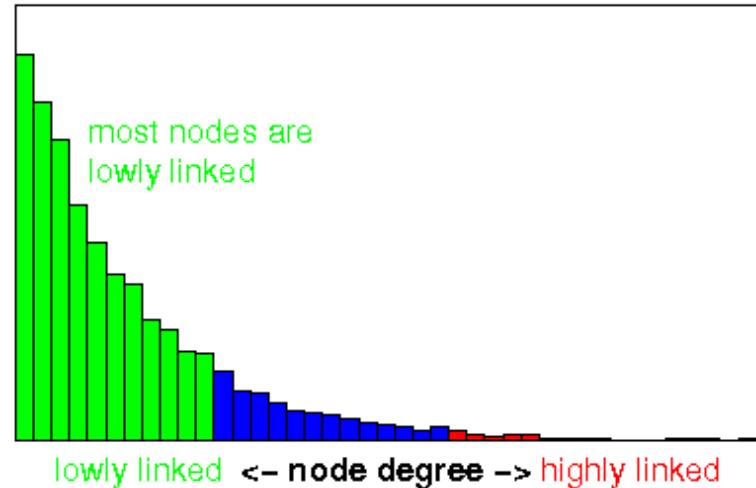
Ricorda: il grado di un nodo corrisponde al numero di vicini di quel nodo

Consideriamo una rete e sia p_d la frazione di nodi che hanno grado d .

Tracciamo la distribuzione dei gradi

In reti “reali”,

- la maggior parte dei nodi ha un grado basso
- esiste una "coda" significativa di nodi con un grado molto più alto



Ricapitolando:Caratteristiche principali delle reti reali

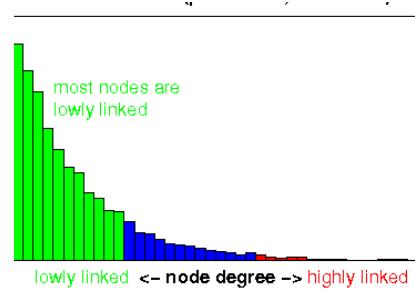
Componente gigante: esistenza di una componente connessa “*molto grande*”

Fenomeno Small-world : distanze “brevi” tra coppie di nodi
(in una stessa componente connessa)

Coefficiente Clustering $\langle C \rangle = \frac{1}{N} \sum_{i=1}^N C_i$ “*elevato*”

$\langle C \rangle$ rappresenta la probabilità che selezionando casualmente due vicini di un qualsiasi nodo, questi hanno un edge che li connette.

Distribuzione dei gradi: con una lunga coda



Network Data-Sets

La ricerca sulle reti su larga scala è stata alimentata in larga misura dalla crescente disponibilità di set di dati dettagliati di reti di grandi dimensioni

- Principali data-sets disponibili in rete su network di grandi dimensioni
 - Grafi di Collaborazioni
 - Wikipedia, Citation graphs
 - Grafi chi-parla-con-chi
 - Microsoft IM, Cell phone graphs
 - Reti di informazioni
 - Hyperlinks
 - Reti tecnologiche
 - Power grids, communication links, Internet
 - Reti naturali e biologiche
 - Food webs, neural interconnections, cell metabolism
- Leskovec's SNAP a Stanford ha un repository di dati su reti di grandi dimensioni
 - <http://snap.stanford.edu/data>

MESSENGER



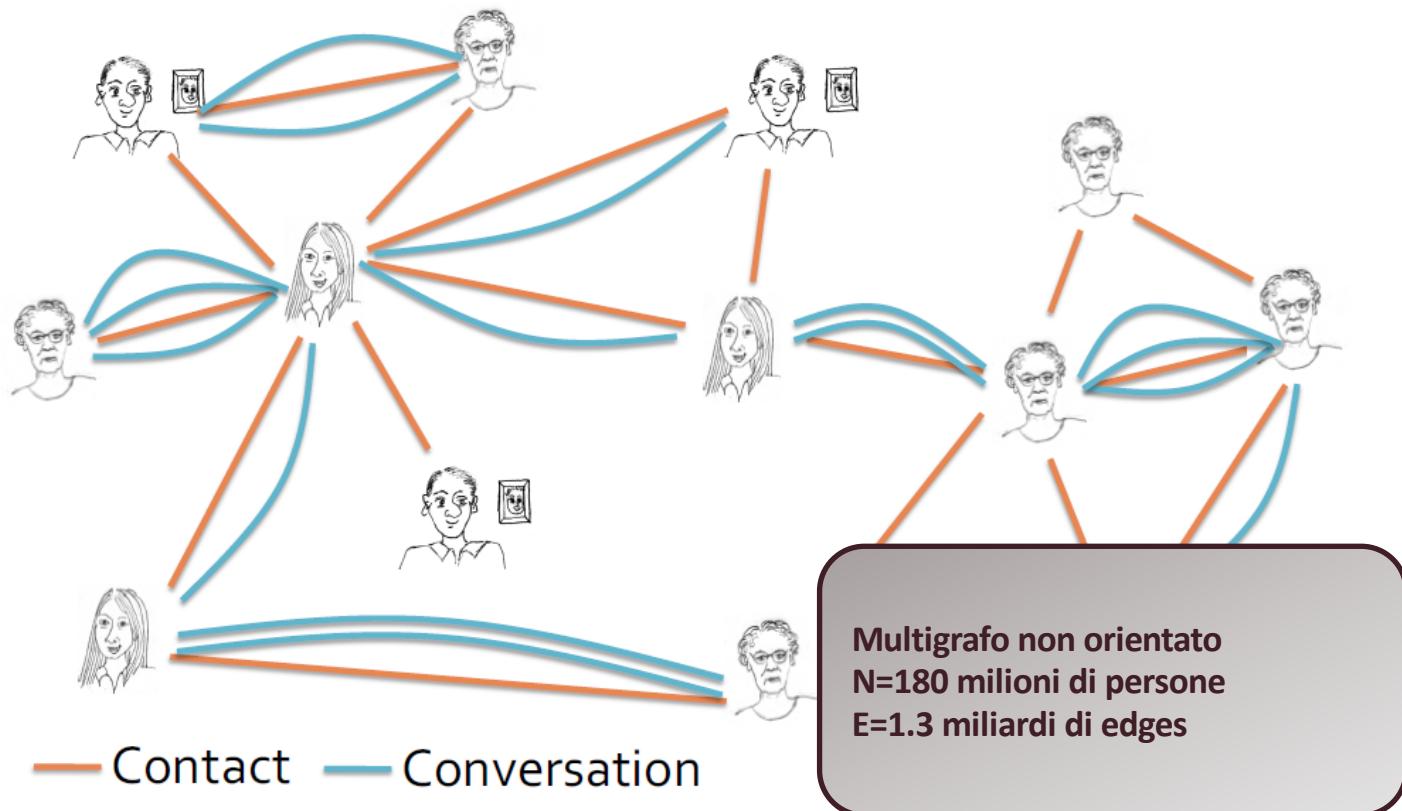
MSN Messenger.

■ 1 month activity

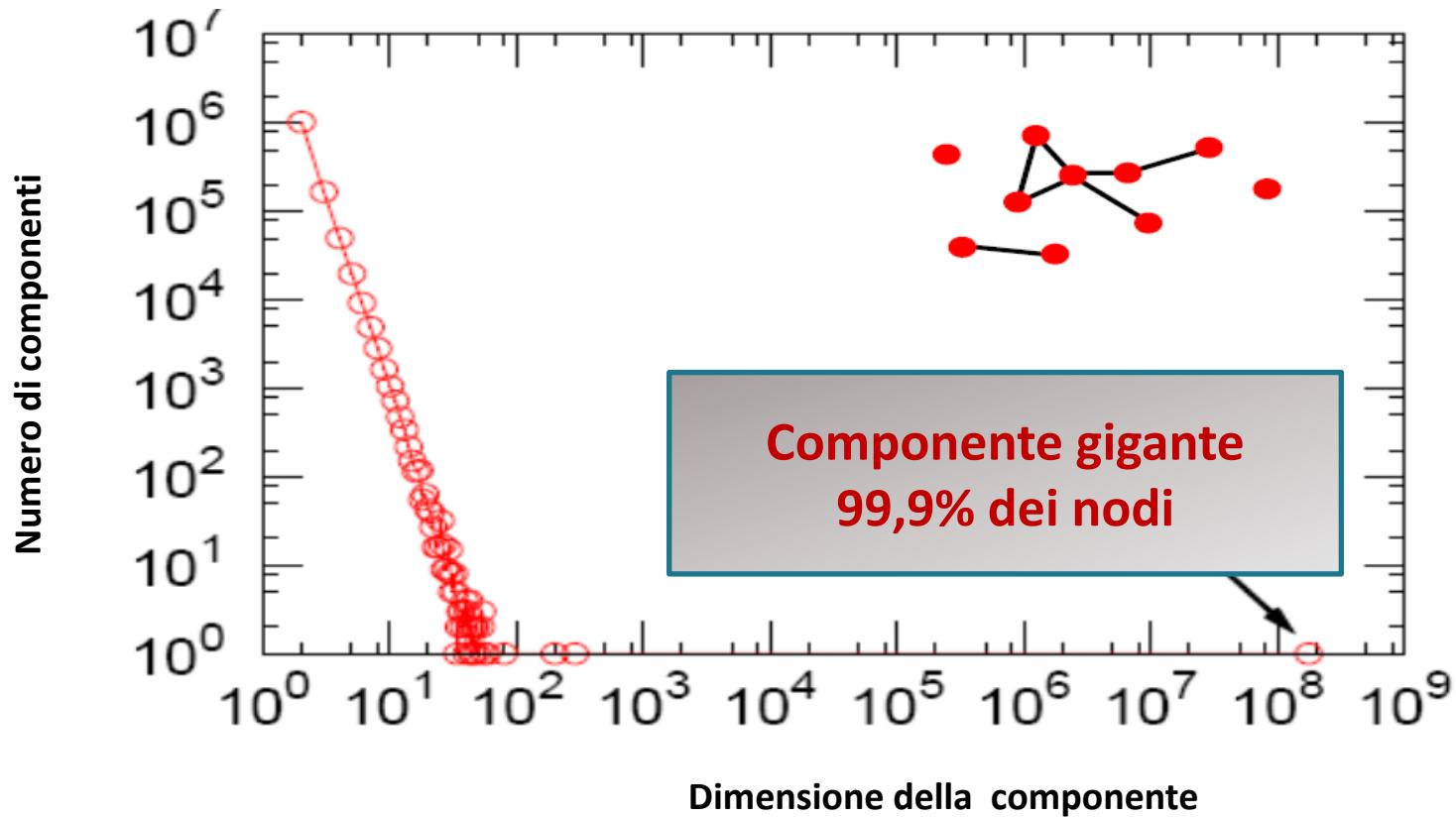
- 245 million users logged in
- 180 million users engaged in conversations
- More than 30 billion conversations
- More than 255 billion exchanged messages

Dati: J. Leskovec, Stanford Univ.

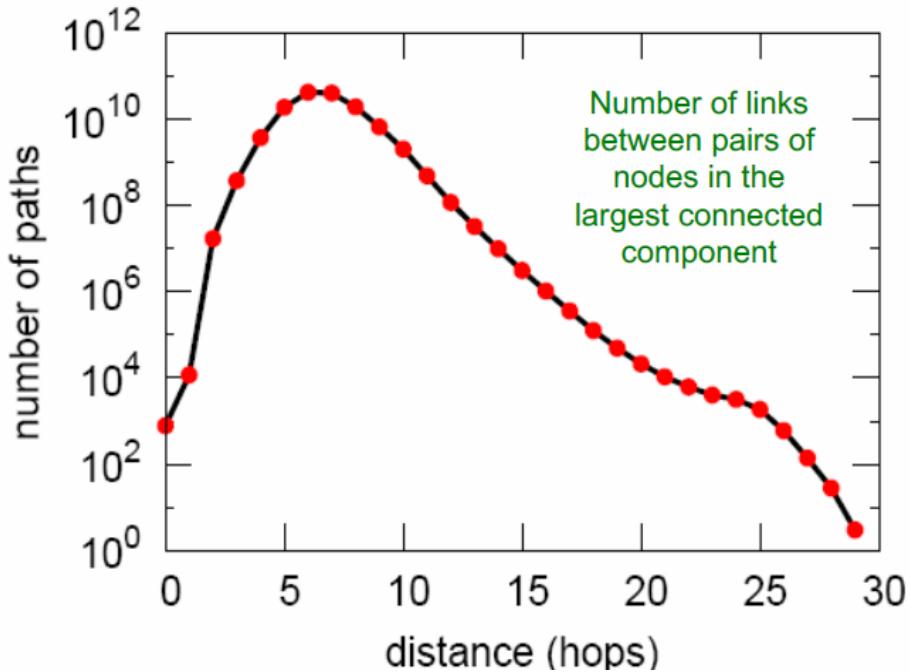
MESSENGER



MESSENGER



MESSENGER



Distanza media 6,6
90% delle coppie a distanza <8

Steps	#Nodes
0	1
1	10
2	78
3	3,96
4	8,648
5	3,299,252
6	28,395,849
7	79,059,497
8	52,995,778
9	10,321,008
10	1,955,007
11	518,410
12	149,945
13	44,616
14	13,740
15	4,476
16	1,542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

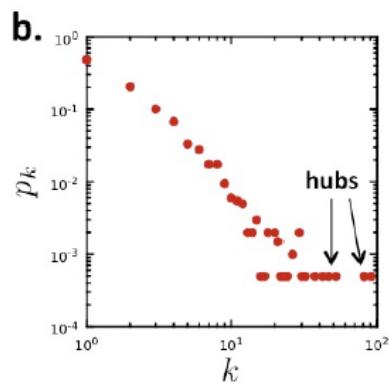
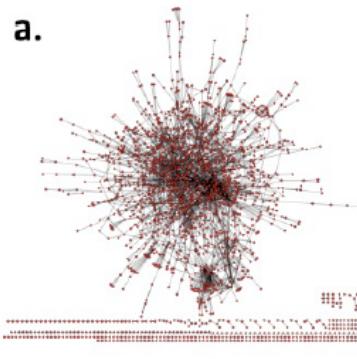
nodes as we do BFS out of a random node

MESSENGER

Ricapitolando

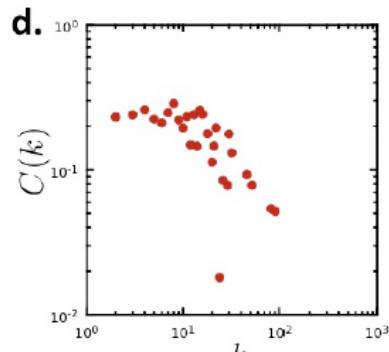
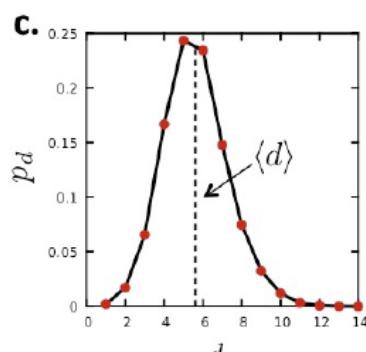
- **Distribuzione dei gradi: molto sbilanciata con grado medio 14,4**
- **Distanza media: 6,6**
- **Coefficiente di clustering 0.11**
- **Componenti: presenza della componente gigante**

PPI (Protein-Protein Interaction) Network



a. Undirected network

N=2,018 proteins as nodes
E=2,930 binding interactions as links.



b. Degree distribution:

Skewed. Average degree $\langle k \rangle = 2.90$

c. Diameter:

Avg. path length = 5.8

d. Clustering:

Avg. clustering = 0.12

Connectivity: 185 components
the largest component 1,647
nodes (81% of nodes)

Ricapitolando

Che cosa vogliamo ottenere dallo studio delle reti?

- Comprendere perchè le reti sono organizzate nel modo in cui sono
- Prevedere il comportamento di sistemi networked

Vedremo (lista non esaustiva)

Modelli

- Erdös-Renyi
- Small-world
- Preferential attachment
- Independent cascade
(modello di diffusione)

Algoritmi

- Community detection:
metodi e modularità
- Ricerca decentralizzata
(cammini brevi)
- PageRank,
- Massimizzazione
dell'influenza

Ricapitolando

Che cosa vogliamo ottenere dallo studio delle reti?

- Comprendere perchè le reti sono organizzate nel modo in cui sono
- Prevedere il comportamento di sistemi networked

Vedremo (lista non esaustiva)

Modelli

- Erdös-Renyi
- Small-world
- Preferential attachment
- Independent cascade
(modello di diffusione)

Algoritmi

- Community detection:
metodi e modularità
- Ricerca decentralizzata
(cammini brevi)
- PageRank,
- Massimizzazione
dell'influenza