Introduzione alla *network analysis*

Raffaele Giammetti (Università Politecnica delle Marche)

1.I network: esempi e applicazioni

La network analysis è un ramo della scienza delle reti che si occupa dello studio delle reti complesse. Al fine di investigare le relazioni complesse, la network analysis adotta teorie e metodi tipici di diversi ambiti di ricerca. Per esempio prende in prestito la teoria dei grafi dalla matematica, la meccanica statistica dalla fisica, il data mining e la visualizzazione delle informazioni dall'informatica, la modellazione inferenziale dalle scienze statistiche e lo studio della struttura sociale dalla sociologia. In sintesi, il Consiglio Nazionale delle Ricerche degli Stati Uniti definisce la scienza delle reti come "*the study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena*".

Ma cosa è un network, e come può essere analizzato? Un network nella sua forma più semplice è una raccolta di punti uniti tra loro in coppie di linee. In termini tecnici i punti sono indicati come vertici o nodi e le linee sono indicate come spigoli, archi o collegamenti. Ad esempio i numeri da 1 a 7 della Figura 1 rappresentano i nodi e le linee che congiungono i numeri sono detti archi.

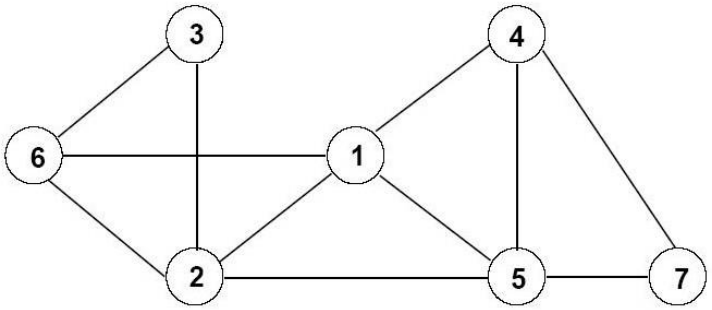


Figura 1. Undirected graph

L'origine della teoria delle reti viene fatta risalire a Eulero e al suo saggio del 1736 in cui è presente una risoluzione matematica dell'enigma dei sette ponti di Königsberg. La città di Königsberg in Prussia (ora Kaliningrad, Russia) era situata su entrambi i lati del fiume Pregel e comprendeva due grandi isole che erano collegate tra di loro e alle due porzioni continentali della città, da sette ponti. Il problema era di escogitare una passeggiata attraverso tutta la città passando per ognuno di quei ponti una sola volta. Eulero formulò il problema in termini di teoria dei grafi, astraendo dalla situazione specifica di Königsberg. Egli rimpiazzò ogni area urbana con un punto, ovvero un vertice o nodo e ogni ponte con un segmento di linea, ovvero uno spigolo, arco o collegamento (Figura 2). Si noti che dai nodi A, B e D partono (e arrivano) tre ponti; dal nodo C, invece, cinque ponti. Questi sono i gradi dei nodi: rispettivamente, 3, 3, 5, 3. Dopo diverse osservazioni e tentativi Eulero formulò il seguente teorema:

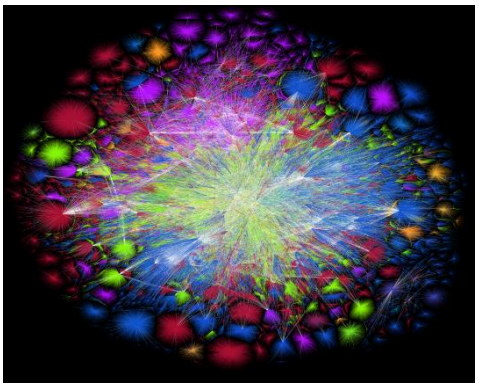
Un qualsiasi grafo è percorribile se e solo se ha tutti i nodi di grado pari, o due di essi sono di grado dispari; per percorrere un grafo "possibile" con due nodi di grado dispari, è necessario partire da uno di essi, e si terminerà sull'altro nodo dispari.

Pertanto è impossibile percorrere Königsberg come richiesto dalla tesi, poiché tutti i nodi sono di grado dispari. Da allora l'utilizzo e studio di grafi si è affinato e sviluppato trovando applicazione in moltissimi ambiti.

Figura 2. Soluzione topologica dell'enigma dei sette ponti di Königsberg.

Le figure di seguito riportano alcuni esempi di network.

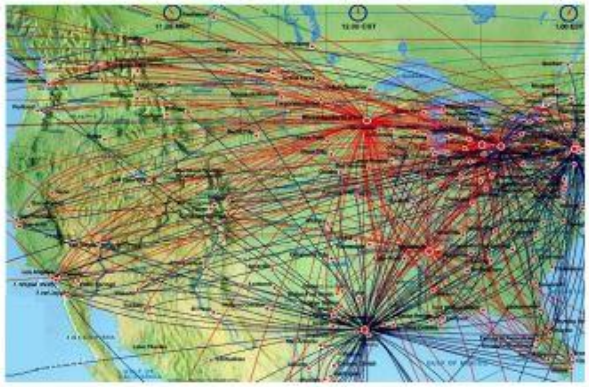
Network tecnologici: Internet, ovvero la rete che connette i computer di tutto il mondo; la rete telefonica; la rete dei trasporti; la rete dei trasporti merci ecc.



a) La struttura di Internet, Opte Project

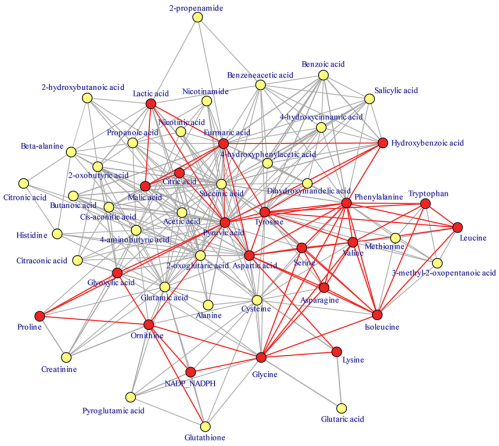
b) Rete telefonica negli USA

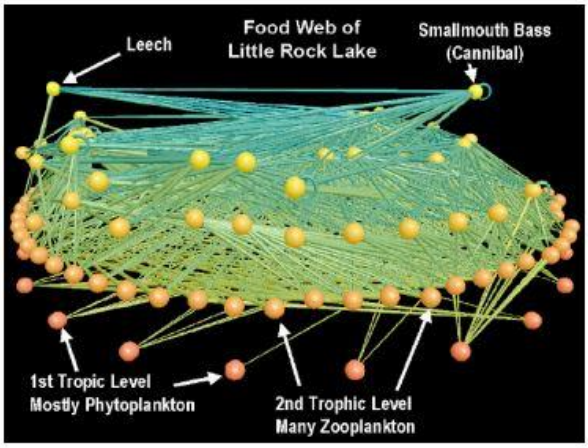
c) Rete trasporti negli USA



d) Rete distribuzione merci negli USA

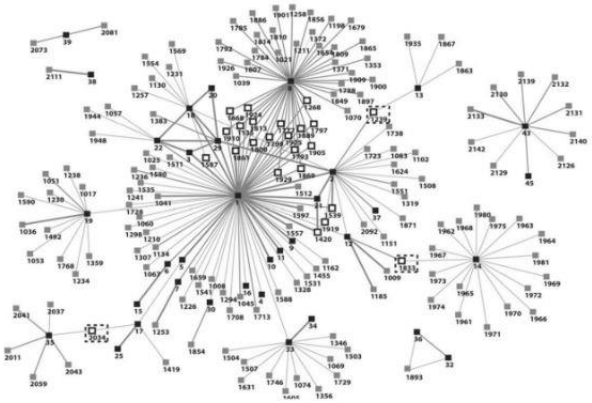
Network biologici, ecologici ed epidemici: le reti metaboliche; le reti alimentari preda-predatore; le reti di diffusione delle malattie ecc.





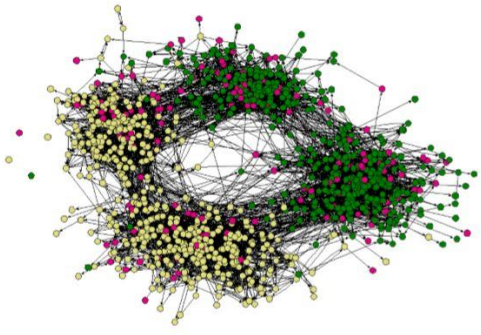
a) Rete metabolica, Delplanke et al. 2018

b) Rete alimentare Little Rock Lake, WI, USA, Yoon et al. 2005

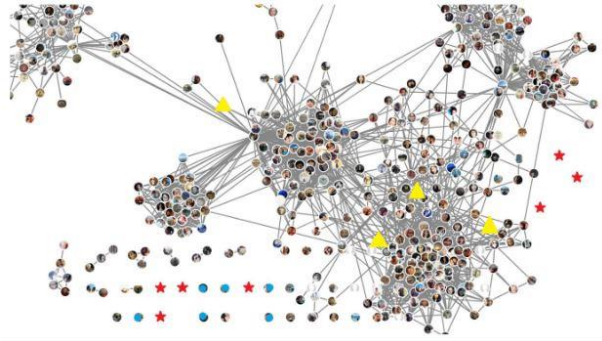


c) Diffusione tubercolosi, McKenzie et al 2007

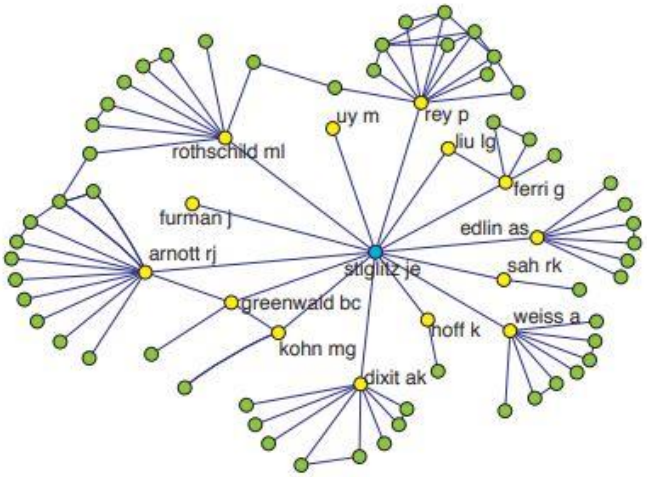
Network sociali: la rete di amicizie; le reti criminali; le reti di collaborazione tra scienziati, attori musicisti ecc,; le communities online ovvero i social network;



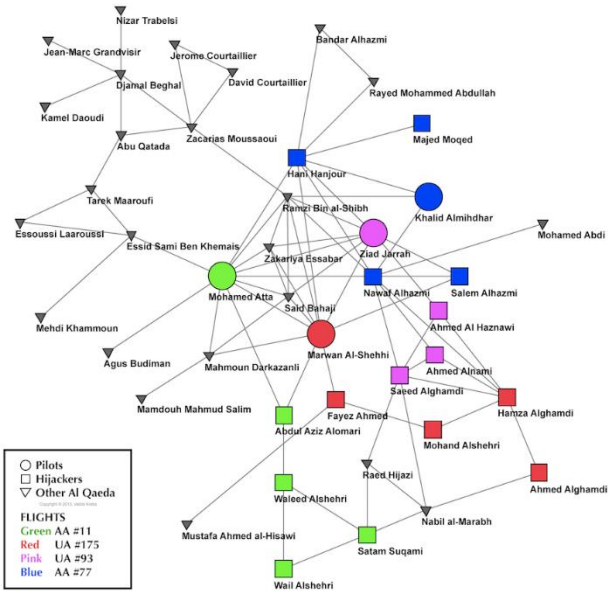
a) Race, school integration, and friendship segregation in America (Moody, 2001)



b) Rete amicizie Facebook



c) Rete collaborazioni J. E. Stiglitz negli anni '90 Goyal 2005



b) Rete terroristi coinvolti nell'attentato dell'unidici settembre 2001, http://www.orgnet.com

2.I network economici

Recentemente la network analysis è stata applicata anche allo studio dell'economia. I network economici studiano sia fenomeni reali che finanziari. Esempi di network reali sono: le reti commerciali, dove i nodi rappresentano i paesi e i link rappresentano relazioni di import/export (Figura 3); le reti produttive dove i nodi rappresentano i settori produttivi di una economia e i link rappresentano relazioni di input/output (Figura 4); le reti di imprese, che sono una versione micro delle reti produttive, ovvero i nodi rappresentano le imprese e i link rappresentano gli scambi tra imprese.

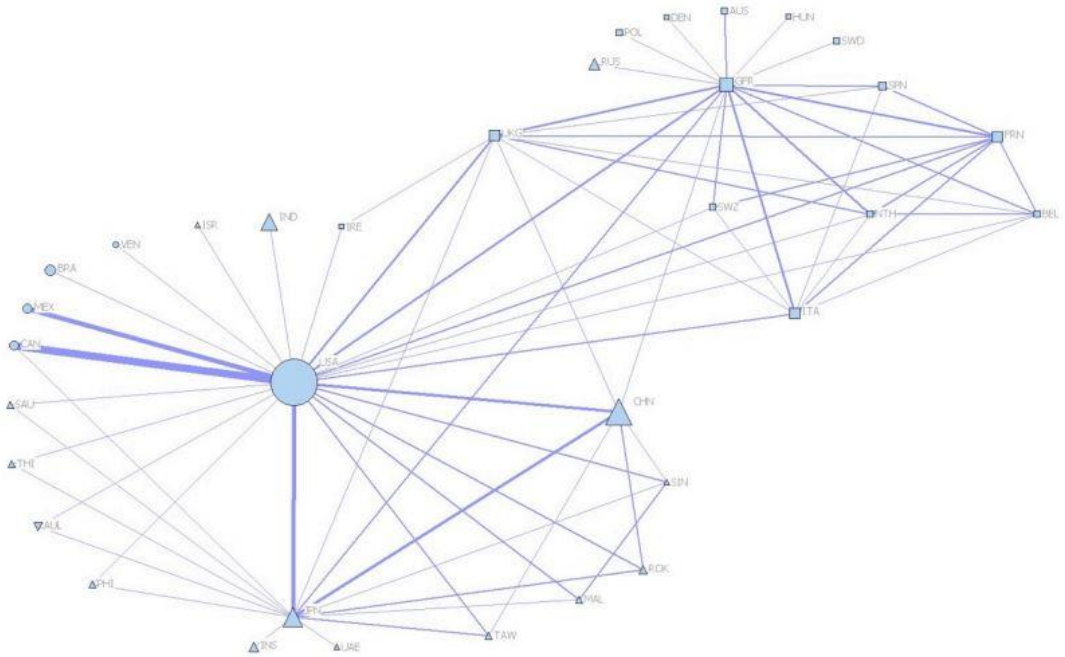


Figura 3. International Trade Network (Fagiolo, 2010).

Lo studio di questi network permette, ad esempio, di verificare il grado di connessione della rete, la distanza tra i nodi e la presenza di cluster o comunità; ancora, studiando i collegamenti, è possibile scovare quali sono i paesi, i settori o le imprese più importanti all'interno della rete e quali sono invece i più periferici. Queste misure, come vedremo, sono alla base della network analysis e possono aiutarci a individuare e spiegare fenomeni complessi come le delocalizzazioni produttive, i fallimenti a catena tra imprese, gli effetti *spillover* ecc.

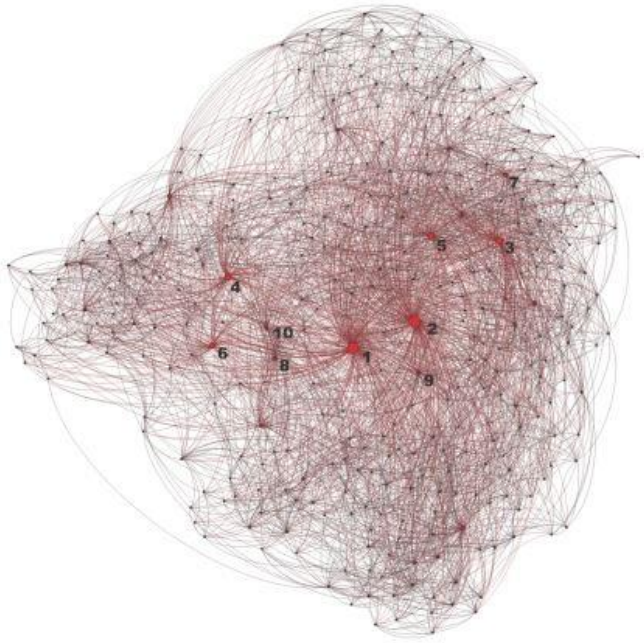


Figura 4. Network produttivo USA costruito con dati input-output (Carvalho, 2014)

Esempi di network finanziari sono: le reti interbancarie, dove i nodi rappresentano le banche e i link rappresentano flussi di liquidità (Figura 5); le reti creditizie dove i nodi possono rappresentare banche, imprese e individui e i link rappresentano relazioni di debito/credito; la rete degli assetti proprietari dove i nodi possono rappresentare istituzioni finanziarie, imprese, investitori privati e paesi e i link rappresentano partecipazioni al capitale azionario (Figura 6).

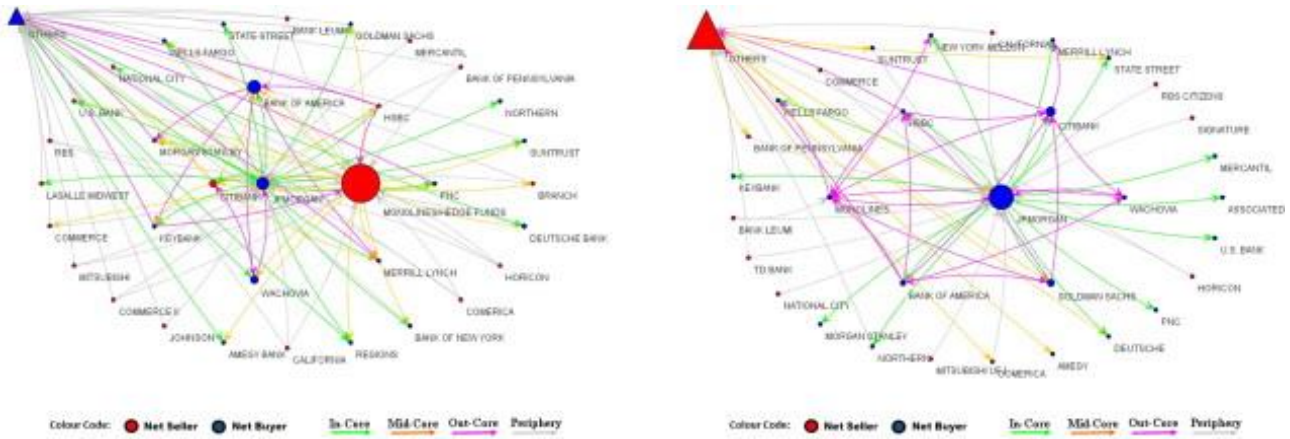


Figura 5. Network interbancari (Markose et al., 2012)

Come per i network reali lo studio di queste reti permette di evidenziare come e quanto i nodi sono connessi e di individuare i nodi più centrali e i più periferici. Individuare il grado di interdipendenza degli agenti di un network finanziario può essere determinante per conoscere e arginare il rischio di contagio in caso di crisi finanziaria. Lo studio dei network finanziari ha infatti spostato l'attenzione dal *too big to fail* al *too interconnected to fail*. A differenza dei precedenti, lo studio a rete degli assetti proprietari è un concetto un po' diverso. In questi network le relazioni tra nodi esprimono rapporti proprietari che oltre una certa soglia implicano il controllo societario. Tali soglie di controllo possono essere raggiunte mediante le partecipazioni indirette che permettono la costruzione di vere e proprie reti di controllo. L'obiettivo nell'analisi di queste strutture proprietarie complesse è di individuare i nodi cui fa capo la rete di controllo e dunque, in ultima istanza, di misurare la concentrazione del controllo del capitale all'interno della rete.

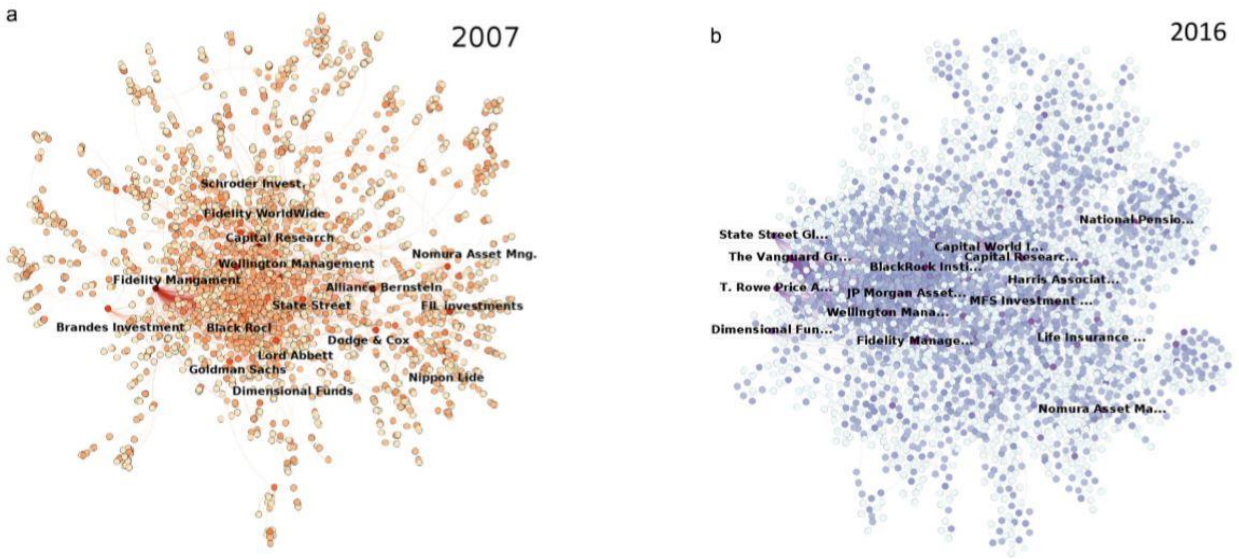


Figura 6. Network degli assetti proprietari del capitale azionario mondiale negli anni 2007-2016 (Brancaccio et al., 2018)

3.Formalizzazione matematica

Un grafo G è una coppia ordinata *G = (N, E)*. *N* è l'insieme di nodi, ed *E* è l'insieme di archi. In alcuni grafi gli archi sono *orientati* e vengono rappresentati da frecce (pensiamo ad esempio alla rete alimentare dove i carnivori mangiano gli erbivori e non viceversa). Un "arco orientato" è dunque un arco caratterizzato da una direzione. In particolare, è composto da una "testa" (rappresentata solitamente dalla punta di una freccia), che si dice raggiunge un vertice in entrata, e una "coda", che lo lascia in uscita. Un "grafo non orientato" è un insieme di vertici e archi dove la connessione *i - j* ha lo stesso significato della connessione *j – i* (la rete di amicizie ad esempio è rappresentata da un grafo non orientato). Ancora, in alcuni grafi gli archi possono essere *orientati* in entrambe le direzioni (pensiamo ad esempio ad una rete di ownership dove una società può essere al contempo partecipata e partecipante).

La struttura di un grafo *G = (N, E)*, può essere rappresentata per mezzo di una matrice detta matrice di adiacenza *A (n, n)* i cui elementi a_{ij} sono 0 se i nodi *i* e *j* non sono connessi e 1 altrimenti. La Figura 7 mostra due semplici grafi e le rispettive matrice di adiacenza. Si noti che la matrice di un grafo orientato (Figura 7, destra) è non simmetrica.

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

Figura 7

Negli esempi di Figura 7 ciascun arco ha la stessa importanza. Infatti, le matrici di adiacenza mostrano che è assegnato il valore di 1 a ciascun collegamento tra i nodi. In molti network questa semplificazione è utile o addirittura essenziale (si pensi ai network sociali). Tuttavia, nella maggior parte delle applicazioni economiche può essere necessario considerare il peso delle relazioni. Definiamo, dunque, grafi *pesati* i grafi rappresentati da matrici di adiacenza A (n, n) i cui elementi sono numeri reali (ad esempio, Figura 8).

$$A = \begin{pmatrix} 0 & 1 & 12 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 8 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

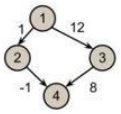


Figura 8

Pertanto, come mostra la Figura 9 possiamo classificare i grafi in *orientati VS non-orientati* e in *binari VS pesati*.

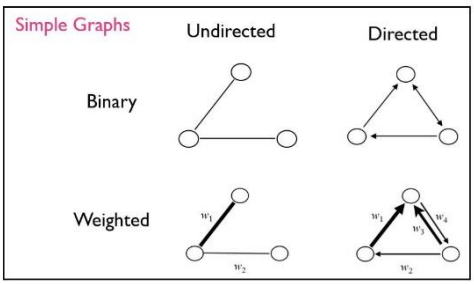


Figura 9

4.Definizioni di base e alcune misure statistiche.

Prima di descrivere le principali misure statistiche dei network, alcune definizioni di base sono necessarie.
Innanzitutto, definiamo *order* il numero *N* di nodi del network e *size* il numero *E* di archi. Guardando la struttura topologica e la connettività del network, definiamo *cammino* da *i* a *z* l'insieme di nodi {i, j, …., z} tale che $i_n\, i_{n+1} \in E$ per ogni n = 0, . . . , z – 1. Ovvero un *cammino* è una sequenza di archi che connettono una sequenza di nodi. Definiamo *percorso* da *i* a *z* l'insieme di nodi {i, j, …., z} $i_m \neq i_n$ per tutti *i,j* tale che $i_n\, i_{n+1} \in E$ per ogni n = 0, . . . , z – 1. Ovvero un *percorso* è un *cammino* in cui un nodo appare al massimo una volta nella sequenza. Definiamo *ciclo* un *cammino* {i, j, …., z} che termina al nodo iniziale e tale che tutti gli altri nodi siano distinti, ovvero tale che {i, j, …., z-1} è un *percorso*.

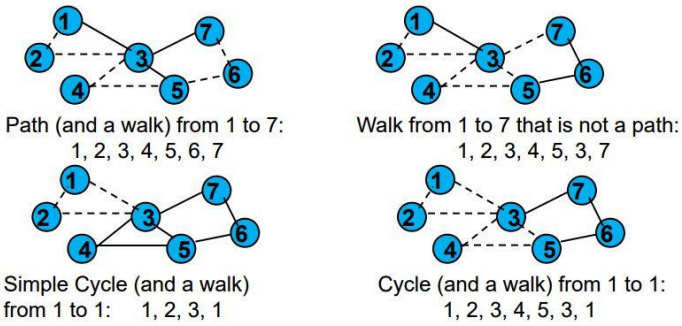


Figura 10

Si definisce *geodesic* o più semplicemente *distanza*, il percorso più breve che collega due nodi i j, ovvero il numero minore di archi che bisogna attraversare per spostarsi da *i* a *j*. Nella Figura 11 $L_{(i,j)}$ è il numero di archi nel percorso più breve tra i vertici *i* e *j* (percorso geodetico). Viceversa la massima distanza tra due nodi del grafo si definisce *diametro*.

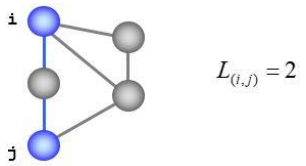


Figura 11

Il fatto che esistano dei percorsi che permettono ad un nodo di raggiungere un altro nodo con il quale non è direttamente collegato è di rilevante importanza in molte applicazioni. L'analisi delle relazioni indirette, infatti, gioca un ruolo cruciale nello studio del contagio, delle strutture proprietarie complesse, nella diffusione delle informazioni e dell'apprendimento o più in generale nella diffusione di diversi comportamenti all'interno di reti sociali. Da un'analisi della struttura topologica e dei percorsi che collegano i differenti nodi del network, è possibile verificare che il network si partiziona naturalmente in sottografi noti come *componenti*. Un network non-orientato si definisce *connesso* (o path-connected) se ogni coppia di nodi è collegata da un percorso. A loro volta, si definiscono

componenti del network i sottografi *connessi*. Un network orientato è *fortemente connesso* se esistono percorsi diretti per ogni coppia di nodi o *debolmente connesso* se i percorsi esistono solo considerando gli archi come fossero non orientati.

Finora abbiamo descritto la rete nel suo complesso giungendo alla descrizione di network più o meno fortemente connessi. A ben vedere, tale misura è studiata anche a livello dei singoli nodi. In particolare il grado di connessione, *degree*, di un nodo *i* è pari al numero degli archi di *i*. In un network orientato distinguiamo tra grado di connessione in entrata, *in-degree*, che è pari al numero di archi che puntano verso il nodo i-esimo, il grado di connessione in uscita, *out-degree*, che è pari al numero di archi che escono dal nodo i-esimo, e il grado di connessione totale, *total degree*, che è pari alla somma tra *in-degree* e *out-degree*. Ad esempio il nodo 4 della Figura 7 ha *degree* 3 nel network non-orientato (sinistra), mentre nel network orientato (destra) il nodo 4 ha un *in-degree* pari ad 1, un *out-degree* pari a 2 e un *total degree* pari a 3. Dividendo il *degree* medio per *n-1* si ottiene una misura della *densità* del network. Quando il network è pesato, si parla più propriamente di *node strength* ovvero la somma degli archi pesati di un nodo. Anche in questo caso vale la distinzione tra network orientati e non-orientati. Quindi, per i network orientati parleremo di *in-strenght*, *out-strenght* e *total strenght* rispettivamente. Ad esempio nella Figura 8 il nodo 2 ha *in-strength* pari a 1, *out-strenght* pari a -1 e *total strenght* 0. Una delle caratteristiche fondamentali di un network è la distribuzione del degree. La distribuzione del degree di un network è la descrizione della frequenza relativa dei nodi che hanno differenti degree. A seconda della distribuzione del degree distinguiamo tra *regular network*, reti in cui ciascun nodo ha lo stesso degree, *random network*, in cui la distribuzione segue una legge esponenziale e *scale-free network*, reti caratterizzati da distribuzioni che seguono una legge di potenza. La Figura 12 mostra un *regular network*. Solitamente si fa riferimento alla mappa delle strade di Manhattan come esempio esplicativo. Un *random network* è un network composto da N nodi in cui ogni coppia di nodi è connessa con probabilità *p*, Figura 13.

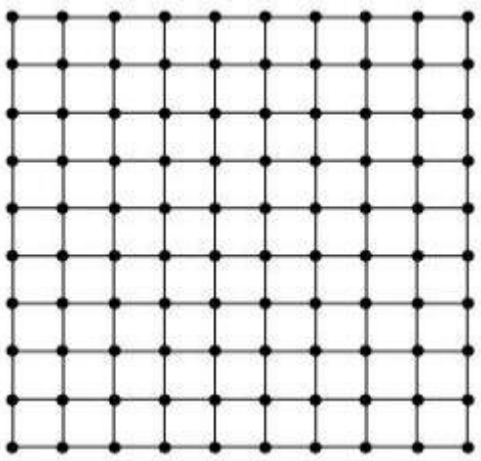


Figura 12. Regular Network

Tuttavia, in molti casi come ad esempio la distribuzione del reddito, l'uso delle parole, la popolazione delle città, la grandezza delle imprese ecc., la distribuzione del degree è asimmetrica e segue una legge di potenza tale che la probabilità *p* di un nodo di possedere un degree *k* è pari a: $p(k) = k^{-\alpha}$, dove $\alpha>1$ è detto fattore moltiplicativo. In tali network, *scale-free*, ci sono pochi nodi altamente connessi e molti nodi poco connessi. A distinguere queste tre macro tipologie di network contribuiscono anche altri due fattori. Il primo è l'*average path length*, ovvero la media dei percorsi tra ogni coppia di nodi. Questa misura ci dice, in media, quanto sono distanti i nodi della rete. Minore è l'*average path length*

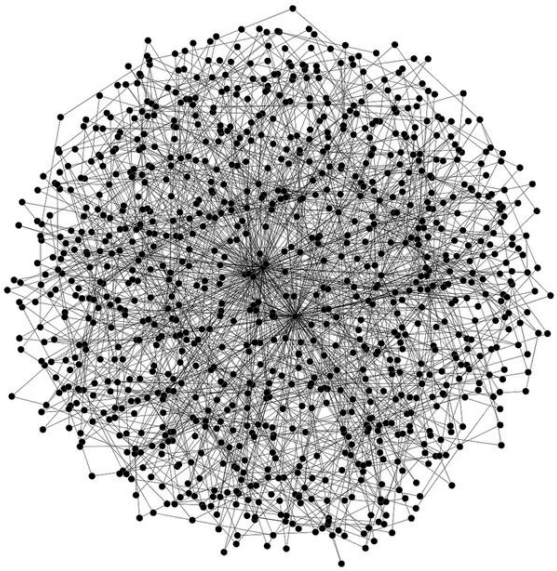


Figura 13. Esempio di Random Network

maggiore è la vicinanza dei nodi e la possibilità di raggiungere un nodo non connesso. Il secondo fattore è il cosiddetto coefficiente di raggruppamento, o *clustering*. In che misura i miei amici sono amici tra di loro? Il coefficiente di *clustering* è una misura del grado in cui i nodi di un network tendono a raggrupparsi insieme. L'evidenza suggerisce che nella maggior parte delle reti del mondo reale, e in particolare nelle reti sociali ed economiche, i nodi tendono a creare gruppi strettamente connessi. In generale il coefficiente di *clustering* di un nodo può essere definito come il numero di triangoli completi che appartengono al nodo, o come il rapporto tra il numero effettivo di collegamenti tra i nodi i, j, k, sul numero totale dei possibili collegamenti. Il coefficiente di *clustering* di un network è la media dei coefficienti di *clustering* dei singoli nodi. Maggiore è il coefficiente di *clustering*, maggiore è il grado di interconnessione del network. Ora, se consideriamo insieme le tre caratteristiche principali di un network, distribuzione del degree, *average path length* e coefficiente di *clustering* possiamo classificare ed individuare le principali differenze tra le tre tipologie di network sopra elencate.

Properties / Networks	Clustering Coefficient	Average Path Length	Degree Distribution
Regular	*High*	High	Equal and fixed In-degrees to each node
Random	Low	*Low*	Exponential/ Poisson
Scale Free/Power Law	Low	Variable	*Fat Tail Distribution*

Tabella 1. Proprietà dei network

Dalla Tabella 1 si evince che: in un *regular network* i nodi hanno lo stesso numero di archi, il coefficiente di *clustering* locale e globale è molto elevato ma non esistono percorsi brevi per raggiungere nodi non connessi direttamente; diversamente i *random network* in cui il degree si distribuisce secondo leggi esponenziali (solitamente distribuzioni di Poisson) mostrano sia un basso *average path length* che un basso grado di raggruppamento; infine gli *scale-free network* hanno una distribuzione dei degree altamente asimmetrica, un *average path length* variabile e un basso coefficiente di *clustering*. A ben vedere, la Tabella 1 mostra anche un'ulteriore tipologia di network. Infatti, gli elementi sulla diagonale sono tipici dei cosiddetti *small-world network* che modellano la maggior parte delle reti sociali ed economiche. Uno *small-world network* è una rete in cui la maggior parte dei nodi non sono vicini l'uno dell'altro (distribuzione dei degree asimmetrica), ma i vicini di un dato nodo sono probabilmente vicini l'uno all'altro (elevato coefficiente di *clustering*) e la maggior parte dei nodi può essere raggiunta da ogni altro nodo da un piccolo numero di passi (basso *average path length*). Il fenomeno *small-world* è stato studiato per la prima volta dal sociologo americano Milgram che nel 1967 ha condotto un esperimento per misurare il numero di intermediari necessari per muovere una lettera da una persona A ad una persona B tramite una catena di conoscenti. Lo psicologo americano selezionò, in modo casuale, un gruppo di statunitensi del Midwest e chiese loro di spedire un pacchetto a un estraneo che abitava nel Massachusetts, a diverse migliaia di chilometri di distanza. Ognuno di essi conosceva il nome del destinatario, il suo impiego e la zona in cui risiedeva, ma non l'indirizzo preciso. Fu quindi chiesto a ciascuno dei partecipanti all'esperimento di spedire il proprio pacchetto a una persona da loro conosciuta, che, a loro giudizio, poteva avere la maggiore probabilità di conoscere il destinatario finale. Quella persona avrebbe fatto lo stesso, e così via, fino a che il pacchetto non fosse stato consegnato al destinatario finale. Milgram si aspettava che il completamento della catena avrebbe richiesto almeno un centinaio di intermediari, rilevando invece che i pacchetti, per giungere al destinatario, richiesero in media solo tra i cinque e i sette passaggi. Da qui l'espressione 'sei gradi di separazione'. Il concetto di mondo piccolo descrive il fatto che la maggior parte dei network reali, nonostante siano di grandi dimensioni, presentano *short path-lengths*, ovvero esiste un percorso relativamente che congiunge due nodi qualsiasi della rete.
La maggior parte delle misure elencate finora sono in predominanza di natura 'macro' nel senso che descrivono caratteristiche generali della struttura dei network. Esistono, tuttavia anche misure 'micro' che permettono di confrontare i singoli nodi del network. Tra queste, rivestono una notevole importanza le misure di *centralità*. Il concetto di *centralità* di un nodo riguarda, in generale, l'importanza di un nodo all'interno della rete. La parola 'importanza' ha un ampio numero di significati e ciò porta alla definizione di diverse misure di *centralità*. Le 4 misure principali di centralità sono:

1- Se per importanza intendiamo il numero di connessioni, allora ci riferiremo alla *degree centrality*. Questa è la misura più semplice di centralità. Un nodo è più centrale quanto maggiore è il numero di connessioni: $c_d = \frac{d_i}{n-1}$
2- Se per importanza intendiamo la vicinanza di un nodo rispetto a tutti i nodi della rete allora ci riferiremo alla *closeness centrality*. In tal caso un nodo è più centrale quanto minore è la distanza media dagli altri nodi: $c_c = \frac{n}{\sum_j d_{ij}}$ dove d_{ij} è la distanza tra il nodo i-esimo e j-esimo.
3- Se per importanza intendiamo la capacità di un nodo di connettere tutti gli altri nodi ci riferiremo alla *betweeness centrality*. In tal caso un nodo è più centrale quanto più questo si trovi su percorsi geodetici (più brevi) tra ogni coppia di nodi della rete. Possiamo pensare al nodo con più alta *betweeness centrality* come il miglior *intermediario* della rete: $c_b = \sum_{h,k} \frac{v_{hk}^i}{g_{hk}}$ dove v_{hk}^i è il numero di percorsi geodetici da *h* a *k* che passano per il nodo i-esimo e g_{hk} è il numero totale di percorsi geodetici tra *h* e *k*.

4- Infine alcune misure di centralità considerano oltre all'importanza di un nodo anche l'importanza dei suoi collegamenti. Queste misure tengono conto della qualità dei collegamenti e si basano sulla premessa che l'importanza di un nodo è determinata dall'importanza dei suoi nodi vicini. Quindi non è rilevante solo quanto un nodo sia connesso o vicino a molti altri nodi, ma piuttosto quanto un nodo sia vicino a molti altri nodi 'importanti'. Questo concetto è alla base dei ranking delle citazioni o del ranking di ordinamento delle pagine di Google. In letteratura si sono affermati diversi indicatori di centralità che tengono conto dell'importanza dei vicini, tra i principali: *Katz centrality, Bonachic centraity, eigenvector centrality, PageRank centrality*.