# Analisi e Visualizzazione di Reti Complesse
# Modulo di Network Science - Parte 1
# Bozza Finale

Lorenzo Bosio

26 agosto 2018

# Indice

# 1 Overview

Lo scorso decennio ha visto un crescente fascino verso la complessa connessione della società moderna. Al cuore di questo interesse c'è l'idea di *rete*, che può essere definita semplicemente come un insieme di elementi connessi. Spesso Facebook, YouTube ed altri siti simili, sono definiti come un social networks, ma in realtà sono più delle piattaforme social media. Rimangono comunque un buon esempio di rete, ma per social network si intende la rete reale, formata dalle relazioni tra persone.

L'evoluzione ha portato a delle reti che trasmettono sempre maggiori moli di dati, i quali ci danno informazioni sulle connessioni e sulla rete stessa; ad esempio ci sono dati che si possono scaricare dai social network per analizzare la propria attività ed i propri collegamenti, oppure dati dati economici e finanziari sugli accordi tra diverse aziende.

## 1.1 Aspects of Networks

Nell'accezione più semplice, una rete è un insieme di oggetti nel quale alcune coppie di questi oggetti sono unite da dei **links**. Questa è una definizione molto flessibile e per questo permette di trovare reti in diversi domini; un esempio è quello studiato negli anni '70 da Wayne Zachary, che fece un esperimento in un club di karate, osservando le persone e collegandole tra loro secondo gli incontri e le relazioni reali; il risultato è illustrato nella figura 1.

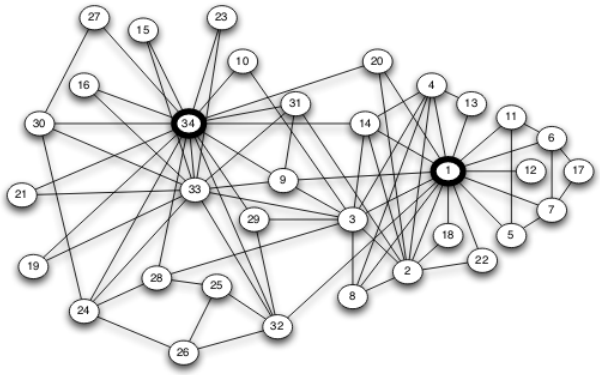



Figura 1: Il club di karate studiato da Zachary.

Con un grafo simile possiamo anche rappresentare i prestiti di denaro tra diverse aziende, i temi di alcuni blogs sulle opinioni politiche, oppure gli scambi di email tra diverse persone, come in figura 2.

Figura 2: Una rete sociale basata sugli scambi di email.

Capiamo quindi che la visualizzazione è importante per rendere l'idea della distribuzione dei dati e dei collegamenti della rete, e quindi per arrivare ad una prima analisi. Ma solamente tramite la visualizzazione non siamo in grado di capire a fondo la rete: abbiamo bisogno di un **linguaggio**, per capire cosa questi dati davvero rappresentano.

### 1.1.1 Behaviour and Dynamics

Un punto di partenza è lo studiare la struttura della rete, in particolare riferendosi alla sua **connectedness**, sia a livello di **struttura**, legami costruiti "forzatamente" da dei vincoli fisici di vicinanza, che a livello di **behavior**, quando si sceglie liberamente con chi legarsi in base alle proprie preferenze. Questo significa che oltre ad un linguaggio abbiamo anche bisogno di un **framework** per ragionare sulle interazione nel contesto della rete, in modo da tenere conto dei comportamenti strategici e dei ragionamenti interni alla rete; queste strategie possono portare cambiamenti nei collegamenti in una rete e per questo si parla di **reti complesse**, che possono evolversi nel tempo.

Ad esempio analizzando il volume di ricerca per YouTube (figura 3) e Flickr (figura 4) nei primi anni 2000 notiamo che questi servizi hanno raggiunto un altissimo livello di popolarità, ma per quale motivo? Si dice "popularity feed itself on its own" e "richer gets richer", per indicare il fatto che la popolarità permette una maggiore diffusione e questa a sua volta aumenta ancora la popolarità. Ma non sempre è questo il caso: ad esempio Facebook ha spodestato come social network il suo predecessore, che al tempo era il più popolare.


Figura 3: La crescita in popolarità di YouTube.


Figura 4: La crescita in popolarità di Flickr.

### 1.1.2 A confluence of Ideas

Con il costante crescere del tema delle reti, della distribuzione dei dati e degli effetti che questo produce, è cresciuto anche l'interesse di diversi campi, come l'economia, la sociologia o l'informatica, verso lo studio di questi fenomeni. Ogni ambito ha portato le proprie idee e concezioni innovative nei suoi studi, soffermandosi sugli aspetti ritenuti più importanti per quel determinato ambito.

## 1.2 Central Themes and Topics

Con queste prime idee, introduciamo ora quelli che saranno i principali oggetti di studio in relazione alle diverse reti, partendo dalla **graph theory**, che si occupa di studiare la struttura della rete, e dalla **game theory**, che offre modelli del comportamento del singolo dove questo è influenzato dal comportamento altrui.

### 1.2.1 Graph Theory

Tornando all'esempio del club di karate di Zachary, sappiamo che alcune persone hanno deciso di rimanere nel club, mentre altri hanno deciso di lasciarlo: si nota che il gruppo che rimane ed il gruppo che lascia hanno molti collegamenti tra membri con la stessa idea e pochi con quelli dell'altro gruppo. Guardando i centri dei due gruppi, notiamo che non sono collegati tra loro: probabilmente questi due membri non andavano d'accordo e gli altri membri hanno deciso chi di appoggiare uno e chi l'altro. Questo fa capire che effetti microscopici possono poi portare ad effetti macroscopici su tutta la rete, da un qualcosa di locale a si ottengono conseguenze osservabili globalmente.

Questo porta l'attenzione sulla **Graph Theory**, in particolare su elementi quali:

- *week and strong ties*, che rappresentano la forza di un collegamento;
- *structural holes*, tra le parti che interagiscono poco con le altre;
- *six degrees of separation*, un fenomeno che indica la distanza media tra due persone nel mondo;
- *structural balance*, un ragionamento su come le spaccature in una rete possono nascere da dinamiche di conflitto a livello locale.

### 1.2.2 Game Theory

La discussione sulla game theory parte dal fatto che spesso in un ambiente gruppi di persone devono scegliere contemporaneamente come comportarsi, sapendo che il risultato dipenderà anche dalle scelte degli altri.Un esempio può essere il dover guidare verso una certa locazione, cercando la strada più corta; se tutti fanno la stessa scelta, ci sarà inevitabilmente traffico in quel percorso. In un sistema complesso si fanno decisioni secondo le proprie necessità ma anche secondo quello che fanno altri: le decisioni di un elemento influenzano anche quelle che prenderanno gli altri, che adottano quindi una propria strategia in conseguenza, volta al miglioramento del proprio **payoff**. Si parla invece di **equilibrio** per indicare uno stato del sistema in cui nessuno ha l'incentivo di cambiare la propria strategia per migliorare il proprio payoff.

### 1.2.3 Markets and Strategic Interaction in Networks

Una volta approfondite graph theory e game theory, possiamo combinarle per produrre dei modelli di comportamento più completi. Questo emerge in

particolare nelle interazioni tra venditori e compratori, dalla cui rappresentazione può emergere l'interessante ruolo dei partecipanti, che possono avere diversi ruoli e posizioni a seconda del numero dei loro collegamenti e della potenza di questi.

### 1.2.4 Information Networks

Anche le informazioni disponibili online hanno una struttura che fondamentalmente rappresenta una rete, basti pensare ai collegamenti tra le pagine del Web ed alla possibilità di muoversi tramite collegamenti. L'uso di questa rete è la base del funzionamento dei motori di ricerca come Google.

Il rapporto tra i creatori di contenuti ed il motore di ricerca è ciò che rende questo tipo di rete dinamico: ogni volta che il motore di ricerca aggiorna il proprio algoritmo, i proprietari delle pagine Web modificano il contenuto di quanto hanno prodotto oppure aggiungono nuove pagine, in modo da poter rimanere tra i risultati con un rank maggiore.

### 1.2.5 Network Dynamics: Population Effects

Se osserviamo per diverso tempo una popolazione, ci rendiamo conto che ciclicamente emergono ed evolvono in essa nuove idee, credenze, opinioni, innovazioni e prodotti. Ci si può riferire a questi come **practices** sociali.

Il modo in cui nascono nuove pratiche dipende direttamente dal fatto che la popolazione **influenza** il comportamento degli altri e dalla tendenza al **conformarsi**. Gli individui possono seguire altri poiché possono pensare che essi abbiano delle loro informazioni aggiuntive, oppure perché al di là del fare la migliore la scelta o meno, l'unione può portare benefici.

### 1.2.6 Network Dynamics: Structural Effects

Non sempre è facile capire come la popolazione possa influenzare il comportamento di altri, ma tener conto della struttura della rete può offrire importanti intuizioni. I meccanismi sottostanti sono presenti sia a livello di popolazione sia a livello locale.

Quando un individuo ha l'incentivo di seguire il comportamento di un altro, può portare ad un effetto a **cascata**. Possiamo fare l'esempio di un racconto grafico giapponese, la cui diffusione si dirama da quattro compratori iniziali.

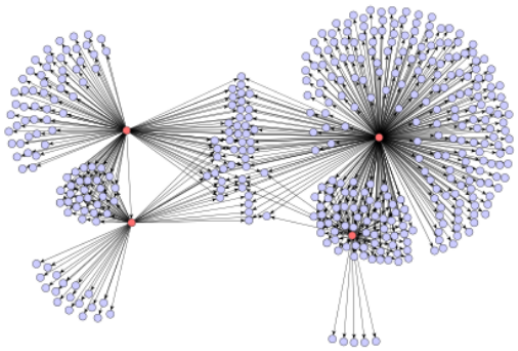


Figura 5: La diffusione del racconto grafico giapponese.

Spesso ci si riferisce a questa diffusione come un **contagio sociale**: in effetti il suo diffondersi è molto simile a quello di una epidemia biologica.

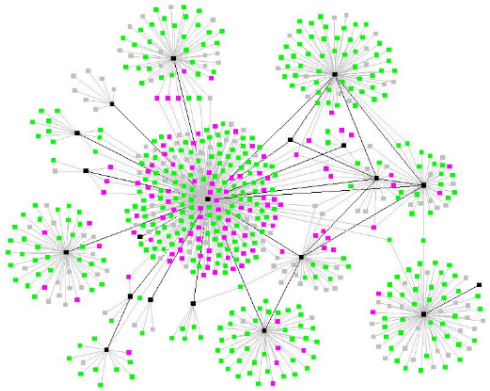


Figura 6: La diffusione di una malattia.

### 1.2.7 Institutions and Aggregate Behavior

Una volta ragionato sulle forze di base che agiscono al di sotto di una rete e del comportamento strategico degli individui, possiamo chiederci come le **istituzioni** designate da una società possano usare queste forze per produrre certi esiti generali; ad esempio è interessante sapere come il mercato finanziario usi le idee ed opinioni delle persone per dare il valore ad un certo oggetto.

Un ruolo importante è dato ai sistemi di votazione, una istituzione che aggrega un comportamento in una popolazione. Inoltre essi possono offrire uno studio sui **prediction markets** e sulle preferenze sociali.

### 1.2.8 Looking Ahead

Tutti i concetti di questo capitolo, accompagnati dalle basi matematiche, motiveranno gli studi più complessi che analizzeranno reti, comportamenti e dinamiche a livello di popolazione.

Il compito è sviluppare la capacità di studiare le reti come sistemi *complessi*, dando importanza alle dinamiche sociali, alle interazioni economiche, alle informazioni online ed ai processi naturali.

# 2 Graphs

In questo capitolo vengono introdotte alcune delle idee di base della teoria dei grafi, in particolare lo studio della struttura della rete.

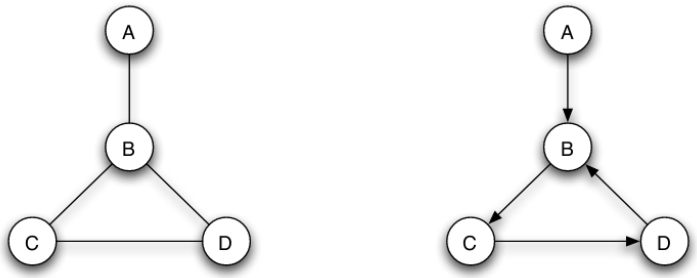
## 2.1 Basic Definitions

Iniziamo con alcune definizioni di base.

### 2.1.1 Graphs: Nodes and Edges

Un **grafo** è una coppia G = (N, E), dove N rappresenta l'insieme dei nodi ed E l'insieme degli archi che connettono i nodi; si può definire come una rete di elementi connessi tra loro. Diciamo che due nodi sono **vicini** se sono connessi da un arco; i vicini di un nodo sono rappresentati dall'insieme di nodi connessi a quel nodo specifico da almeno un arco.

La relazione tra due nodi rappresentata da un arco può essere simmetrica o asimmetrica; per poter rappresentare queste due tipologie definiamo rispettivamente i grafi **non orientati** ed i grafi **orientati**. Nei primi gli archi non hanno direzioni e sono rappresentati come una semplice linea, mentre negli ultimi gli archi sono orientati e rappresentati da delle frecce; inoltre in questi grafi i vicini vengono divisi in vicini **in ingresso** ed **in uscita**.




(a) *A graph on 4 nodes.* (b) *A directed graph on 4 nodes.*

Figura 7: Due grafi: (a) uno non orientato, ed (b) uno orientato.

### 2.1.2 Graphs as Models of Networks

I grafi sono utili perché possono rappresentare un **modello matematico** di una struttura di rete. Possiamo citare l'esempio di ARPANET, il predecessore di Internet, che inizialmente era composto da tredici località principali degli USA connesse tra loro, come in figura 8.

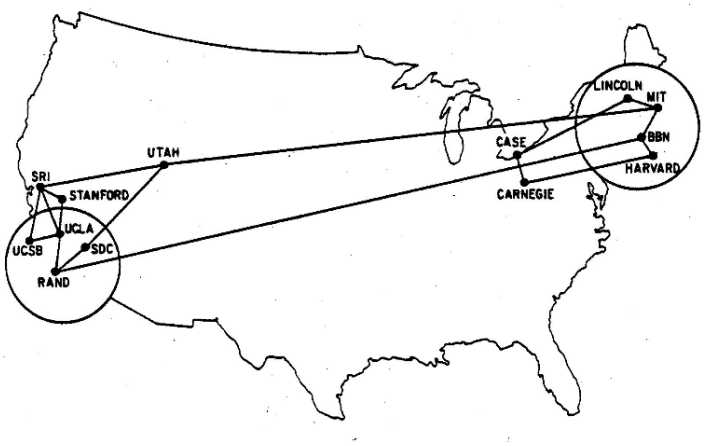



Figura 8: La rete definita da ARPANET.

Astraendo la mappa geografica che rappresenta la rete, otteniamo un grafo come quelli descritti in precedenza, mostrato nella figura 9.

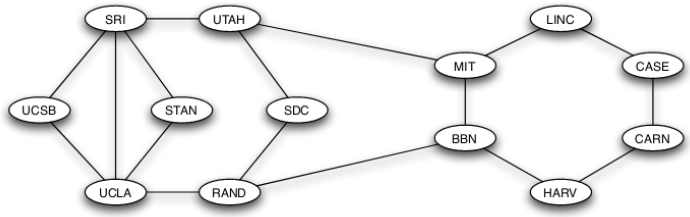
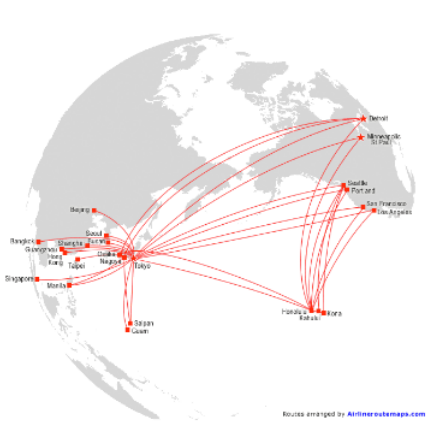



Figura 9: Il grafo ottenuto astraendo la mappa di ARPANET.

I modelli matematici servono ad effettuare una rappresentazione astratta e più semplice di un sistema o di una rete più complesso dell'originale; a questa rappresentazione possiamo inoltre aggiungere altre proprietà, come ad esempio il peso degli archi. I modelli matematici vengono usati per trovare proprietà o pattern dal dominio: da domini diversi possiamo avere pattern simili, che permettono di esprimere delle leggi universali, indipendenti quindi dal dominio analizzato.

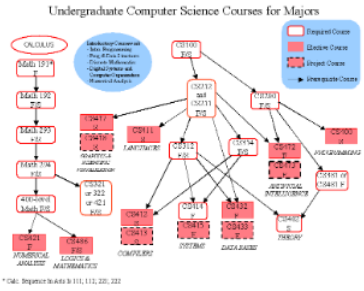
Per fare alcuni esempi vediamo che evidenziando in una mappa le tratte aeree dei principali aeroporti del mondo, queste formano un grafo; osservando la mappa di una metropolitana notiamo che anch'essa legata ai trasporti, ha caratteristiche simili a quella vista in precedenza: si parla quindi di *transportation networks* per definire questo tipo di sistema. Anche il filesystem di Linux può essere considerato come una rete per le sue numerose dipendenze, definendo la tipologia delle **dependency networks**. Un ultimo esempio può

essere quello di un ponte, in cui i giunti rappresentano i nodi ed i collegamenti fisici gli archi; in questo caso si parla di **structural networks**.



(a) *Airline routes*


(b) *Subway map*




(c) *Flowchart of college courses*

(d) *Tank Street Bridge in Brisbane*

Figura 10: Alcuni esempi di grafi.

## 2.2 Paths and Connectivity

Proseguiamo lo studio dei grafi introducendo altri concetti.

### 2.2.1 Paths

È importante l'idea che spesso le cose possono muoversi e viaggiare attraverso gli archi di un grafo, toccando diversi nodi in sequenza: questo può essere il caso di un passeggero che prende una serie di voli facendo diversi scali, oppure una sequenza di pagine visitate durante la navigazione in Internet.

Questo porta a definire il concetto di **cammino**, rappresentato semplicemente come una sequenza $N_1, N_2, N_3, \ldots, N_n$ di nodi con la proprietà che

ogni coppia consecutiva nella sequenza è collegata da un arco. In un cammino possiamo anche trovare nodi ripetuti, ma nel caso in cui non ce ne siano si parla di cammino **semplice**.

### 2.2.2 Cycles

Un caso particolare di cammino non semplice è quello del **ciclo**, in cui il primo e l'ultimo nodo coincidono, in una struttura ad anello. Spesso nelle communication e nelle transportation networks sono presenti dei cicli per questioni di ridondanza, per poter offrire una strada secondaria in caso di problemi o strategie particolari.

### 2.2.3 Connectivity

Si parla di **grafo connesso** quando esiste sempre un cammino tra due nodi qualsiasi del grafo. Non c'è una ragione per pensare che tutti i grafi (e le reti che questi rappresentano) siano connessi, ma ci sono regole matematiche ed assunzioni che possono portarci a formulare delle proprietà anche nei grafi disconnessi.

### 2.2.4 Components

Quando un grafo non è connesso è naturale che venga diviso in un insieme di "pezzi" connessi, ovvero dei gruppi di nodi connessi quando considerati singolarmente e senza sovrapposizioni. Questi gruppi sono chiamati **componenti connesse** e godono di queste proprietà:

1. ogni nodo nella componente ha un cammino verso ogni altro nodo nella componente;

2. non ci sono sottoinsiemi più grandi in cui la proprietà 1 è valida, ovvero la componente non è parte di una componente più grande.

Dividere un grafo in componenti da analizzare è uno dei metodi principali per descriverne la struttura.

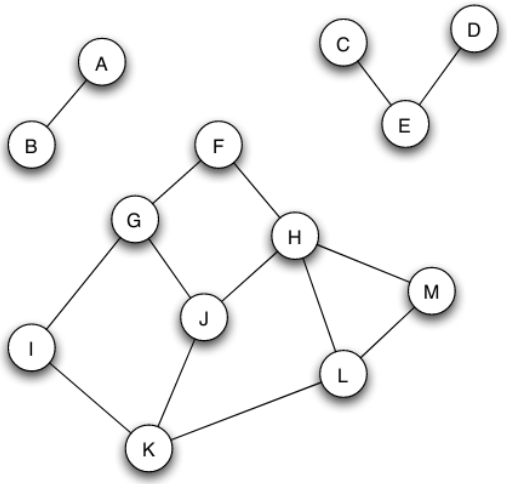



Figura 11: Un grafo con tre componenti connesse.

### 2.2.5 Giant Components

Definiamo le **giant components** come componenti connesse che contengono una parte significante di tutta la rete; in generale, quando una rete possiede una componente gigante ne contiene solamente una, e si può considerare solo questa parte del grafo, tralasciando le componenti minori, per analizzare le proprietà principali.

La nozione di componente gigante è anche utile per ragionare su una rete su una scala minore: ad esempio La figura 12mostra le relazioni in una scuola americana in un periodo di 18 mesi. Questa informazione viene usata per studiare la diffusione delle malattie sessualmente trasmissibili.

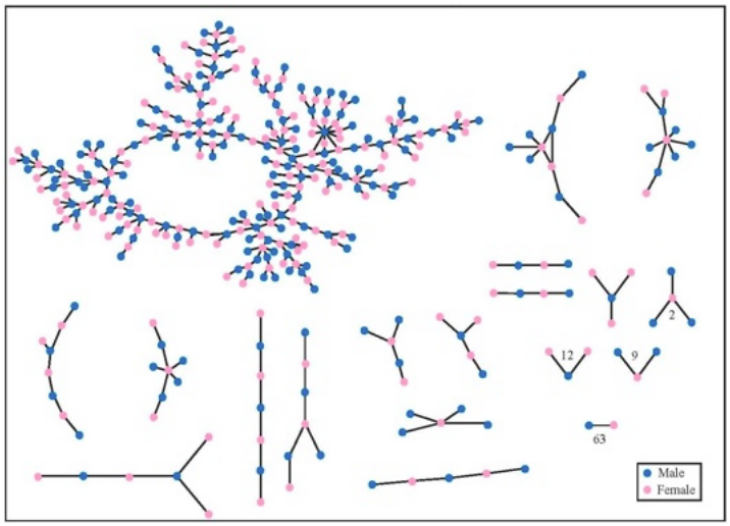



Figura 12: Un grafo in cui i nodi rappresentano gli individui e gli archi le relazioni tra essi.

## 2.3 Distance and Breadth-First Search

Spesso è interessante non solamente chiedersi se due nodi siano connessi da un cammino, ma anche sapere quanto questo cammino sia lungo. Definiamo quindi la **lunghezza** di un cammino, che indica il numero di step che contiene dall'inizio alla fine, ovvero il numero di archi da cui il cammino stesso è formato. Chiamiamo **distanza** la lunghezza del cammino più corto da tra due nodi.

### 2.3.1 Breadth-First Search

Per i grafi più piccoli è facile trovare la distanza tra due nodi qualsiasi semplicemente guardando la sua rappresentazione visiva, ma se le dimensioni del grafo aumentano questo procedimento diventa sempre più difficile; serve quindi un algoritmo che possa trovare sistematicamente la distanza tra qualsiasi coppia di nodi.

I passi da seguire per calcolare le distanze da un nodo di partenza sono i seguenti:

1. dichiarare che tutti i vicini siano a distanza 1;
2. trovare tutti i loro vicini, escludendo quelli già presenti al punto 1, e dichiararli a distanza 2;

3. trovare tutti i vicini dei nodi al punto 2, escludendo quelli trovati in precedenza, e dichiararli a distanza 3;

4. continuare con questo processo finché non si sono esplorati tutti i nodi.

Questo processo è noto come **Breadth First Search**, l'algoritmo più efficiente per trovare le distanze in un grafo.

### 2.3.2 The Small-World Phenomenon

Tornando al discorso delle componenti connesse in un grafo, possiamo ora aggiungere delle proprietà qualitative a questo aspetto. Analizzando un'amicizia con una persona straniera, sono tre semplici collegamenti, ovvero quello con l'amico, che a sua volta è connesso coi propri genitori, che a loro volta hanno degli amici, che portano ad una realtà completamente distante e diversa che non ha nulla a che fare con se stessi. Questa idea che il cammino sia più corto di quanto ci si aspetti è formalizzata dal concetto di **small-world phenomenon**.

Un altro modo di esprimere questa idea è quello dei **sei gradi di separazione**, ossia l'idea che in media ci sono sei archi che ci separano da ogni altra persona del mondo in un ipotetico grafo. Questa convinzione è inizialmente nata da Stanley Milgram, che fece un esperimento, scegliendo 296 persone casuali in Nebraska e mandando loro una lettera chiedendo di cercare di spedirla un certo agente di cambio a Boston, affidandosi alle loro conoscenze. Ogni lettera conteneva la lista di chi l'aveva ricevuta, e con 64 catene terminate calcolò la distanza media di sei persone, da cui viene il detto "six worlds apart".

### 2.3.3 Instant Messaging, Paul Erdös e Kevin Bacon

L'esperimento di Milgran aveva la limitazione di essere limitato ad un certo ambito geografico e di non avere dati ben certi su cui basarsi. Per questo anni dopo fu condotto un altro esperimento chiamato "six degrees of messaging" nell'ambito dell'instant messaging, dove i dati sono sicuri e memorizzati all'interno di un calcolatore. Nel 2008 Leskovec e Horvitz della Microsoft hanno analizzato i dati a loro disposizione, usando 240 milioni di nodi e studiando 1000 utenti casuali, ottenendo una lunghezza media di 6.6 tra i cammini che li collegano. Una variante di questo studio fu effettuata anche da Facebook, riscontrando una media di lunghezza del cammino tra 4.4 e 5.7, ottenendo quindi un collegamento più breve dei precedenti.

È anche doveroso citare Paul Erdös, uno dei più grandi studiosi in ambito matematico, dai cui lavori si è ottenuta una *collaboration network* che rap-

presenta tutti i lavori di altri matematici in collaborazione con Erdös stesso. La distanza di un matematico con il nodo principale di questa rete è nota come "numero di Erdös". Un esempio simile è quello de "l'oracolo di Kevin Bacon", una rete di collaborazione stavolta basata sui film in cui l'attore ha recitato. Il "numero di Bacon" è la distanza tra un qualsiasi attore e Bacon stesso in questo grafo.

## 2.4 Network Data Sets: An Overview

L'esplosione della ricerca su reti di larga scala degli ultimi anni è dovuta principalmente alla nuova disponibilità di grandi insieme di dati e la informazioni che questi possono dare a seconda della loro struttura e provenienza. Diamo quindi una panoramica dei principali **Network Data Sets**.

### 2.4.1 Collaboration Graphs

I **collaboration graphs** registrano le relazioni in un particolare contesto, come ad esempio le apparizioni nello stesso film, le pubblicazioni di articoli con altri autori, gli autori delle pagine di Wikipedia, i partecipanti di un raid o altre attività su WoW.

Spesso i collaboration graphs sono usati per studiare il dominio specifico su cui sono basati; ad esempio i sociologi che studiano l'ambito economico sono interessati alle relazioni tra le diverse aziende a livello di direttivo, quelli che invece studiano il contesto delle ricerche scientifiche sono interessati a trovare delle relazioni tra autori addirittura di secoli diversi.

### 2.4.2 Who-Talks-to-Whom Graphs

Un esempio di **who-talks-to-whom graph** è quello usato nell'esperimento della Microsoft, in cui sono registrate le conversazioni di IM; possono essere citati anche i log degli scambi di email tra diverse persone oppure dei grafi delle chiamate con archi pesati dalla durata della comunicazione (call graphs).

In tutti questi tipi di dati i nodi rappresentano clienti, impiegati o studenti dell'organizzazione che mantiene i dati, ed essi in generale tengono alla propria privacy, senza apprezzare quanto facilmente si possano ricostruire le loro attività da questi dati. Quello della privacy è infatti uno degli argomenti di maggiore interesse negli ultimi tempi per questi studi.

Un altro particolare ambito della relazione who-talks-to-whom è quello chiamato **who-transact-with-whom**, che rappresenta la struttura del mercato finanziario e analizza la relazione tra l'accesso al mercato e la potenza sul mercato stesso in termini di prezzi.

### 2.4.3 Information Linkage Graphs

Gli snapshot del Web sono esempi di insiemi di dati di rete, in cui i nodi sono le pagine Web e gli archi orientati rappresentano i collegamenti tra esse. Questi dati sono importanti sia per le dimensioni che per le informazioni che contengono e che possono rivelare a livello sociale ed economico, dai blog personale alle pagine commerciali delle aziende.

Lavorare con questa quantità di dati è difficile, a particolare dal manipolare correttamente i dati stessi. Per questo motivo molte ricerche sono state effettuare su un sottoinsieme ridotto e ben definito di blog, pagine di Wikipedia, social networks e recensioni su siti di shopping.

Questi studi sono i principali in relazione al Web, in particolare il ramo della "citation analysis" che a partire dal ventesimo secolo ha analizzato la struttura della rete delle citazioni tra papers scientifici per seguire l'evoluzione della scienza.

### 2.4.4 Technological Networks

Sebbene il Web sia costruito su una complessa tecnologia, sarebbe sbagliato pensarlo come una rete tecnologica: esso è infatti una proiezione delle idee, informazioni e strutture sociale ed economiche create dagli umani.

Anche le reti fisiche sono quindi riconducibili a reti economiche, che rappresentano le interazioni tra organizzazioni e compagnie. Su Internet queste relazioni sono osservabili grazie ad una particolare visione della rete a due livelli. A livello più basso i nodi sono i router ed i computer ed un arco indica un collegamento fisico tra di essi. A livello più alto invece questi nodi sono raggruppati in "sistemi autonomi", ognuno controllato da un diverso fornitore del servizio. Questi sistemi autonomi possono essere analizzati attraverso una rete di tipo who-talks-to-whom che rappresenta gli accordi di trasferimento dati tra i fornitori dei servizi.

### 2.4.5 Networks in the Natural World

La struttura dei grafi può anche essere legata alla biologia ed altre scienze naturali. Possiamo fare tre esempi principali di questo tipo di rete:

1. le "food webs" rappresentano una relazione di chi-mangia-chi, rappresentata da n arco orientato, tra diverse specie, indicate dai nodi del grafo. Studiare questo tipo di rete può portare a capire le cause dell'estinzione delle diverse specie;

2. le reti usate per studiare le interconnessioni tra i neuroni nel cervello di un individuo, dove i nodi rappresentano i neuroni ed gli archi una

connessione. Questi studi portano a capire i legami tra specifici moduli nel cervello e come essi si relazionano tra loro;

3. la reti complesse usate per rappresentare il metabolismo di una cellula, nelle quali i nodi rappresentano i composti principali e gli archi le reazioni chimiche che avvengono tra loro. I ricercatori sperano che questi studi possano portare luce sulle complesse reazioni che avvengono all'interno delle cellule e suggerire dei metodi per contrastare gli organismi patogeni che disturbano il metabolismo in maniere specifiche.

# 3 Strong and Weak Ties

Uno dei più potenti ruoli di una rete è quello di legare proprietà locali con proprietà globali, in modo da evidenziare come semplici processi a livello di singoli nodi possano avere effetti complessi su una intera popolazione.

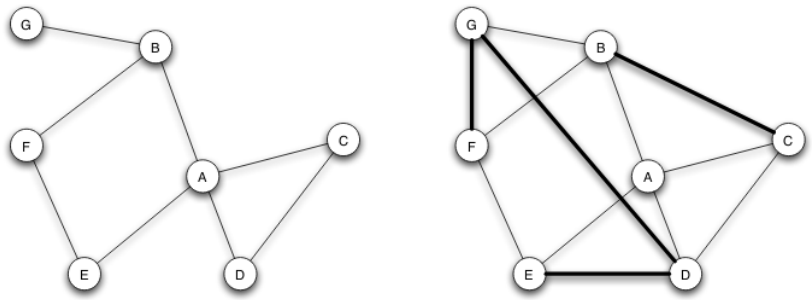
Negli anni '60 Mark Granovetter condusse un esperimento come tesi di dottorato, intervistando diverse persone e chiedendo loro come avessero trovato il loro nuovo lavoro. Ciò che emerse fu che la maggior parte degli intervistati aveva trovato lavoro attraverso un contatto, ma che questo contatto era spesso solamente un conoscente e non un amico. L'ipotesi di Granovetter per spiegare questo fenomeno era che ci fosse una ragione strutturale, ovvero come queste relazioni siano situate in diverse parti della rete, ed una interpersonale, ovvero il concentrarsi sulle conseguenze locali di una amicizia forte o debole; queste due motivazioni trascendono dal particolare ambito della ricerca del lavoro ed offrono spunti sull'architettura generale della rete sociale.

## 3.1 Triadic Closure

Fino ad ora abbiamo trattato le reti come strutture statiche, ma non sempre è così, anzi succede spesso che dei nodi entrino o lascino la rete nel tempo. Uno dei motivi per cui una rete si può evolvere nel tempo fu formulato da Rapoport nel 1953. Egli stabilì la proprietà di **triadic closure**, affermando che:

> Se due nodi hanno una connessione con un nodo in comune, probabilmente nel futuro anche questi due saranno connessi tra loro.

Le nuove connessioni potranno portare a nuove connessioni tra nodi, sempre più distanti da ciò che ha scatenato l'evoluzione, come mostrato in figura 13.




(a) *Before new edges form.* (b) *After new edges form.*

Figura 13: Gli effetti della triadic closure nel tempo.

### 3.1.1 The Clustering Coefficient

La formulazione della proprietà di triadic closure ha portato alla definizione di una semplice misura per catturarne la presenza. È stato quindi introdotto il **clustering coefficient**, che indica la probabilità che i vicini di un nodo siano connessi tra loro ed è calcolata come il numero di connessioni tra i vicini del nodo specifico diviso per il numero totale delle possibili connessioni tra i vicini. Più forte opera la triadic closure, più alto è il coefficiente di clustering tenderà ad essere, portando i triangoli delle relazione a "chiudersi".

È importante segnalare che un alto clustering coefficient è molto frequente nelle social networks, a causa della transitività delle amicizie. Se invece si genera una rete con lo stesso numero di nodi ma con gli archi in ugual numero ma assegnati a caso, si ottiene una rete con un clustering coefficient molto minore.

### 3.1.2 Reasons for Triadic Closure

La triadic closure è un concetto abbastanza naturale ed è facile trovarne esempi nella propria esperienza che possono portare a dedurne le cause principali:

- *opportunità*: se due nodi hanno un amico in comune, è probabile che avranno l'opportunità di incontrarsi;
- *fiducia*: il fatto che due nodi abbiano un amico in comune pone la base per fidarsi l'uno dell'altro;
- *incentivo*: un nodo potrebbe raggrupparne altri due, essendo amico di entrambi.

Questi tre motivi combinati portano idealmente alla "chiusura" delle relazioni triangolari in un grafo.

L'importanza di questa proprietà è dimostrata attraverso diversi studi, anche nell'ambito della psicologia sociale; un esempio è stato riscontrato da Bearman e Moody, che scoprirono che le teenager con un basso clustering coefficient tra le proprie amicizie sono più portate a contemplare il suicidio.

## 3.2 The Strength of Weak Ties

Tornando all'esempio di Granovetter, possiamo ora capire come la triadic closure sia uno dei motivi principali che lo portarono a formulare la sua ipotesi.

### 3.2.1 Bridges and Local Bridges

Spesso le informazioni riguardanti un buon lavoro sono scarse e sentirle da parte di qualcun altro implica che egli abbia accesso ad informazioni a noi sconosciute. Questo porta a pensare che ci siano amicizie "particolari", anche all'interno di un grafo. In figura 14 vediamo che B è amico di A, ma non fa parte della cerchia degli amici più stretti. L'arco che lega A e B è particolare e viene chiamato **bridge** perché se rimosso separa i due nodi in componenti diverse.

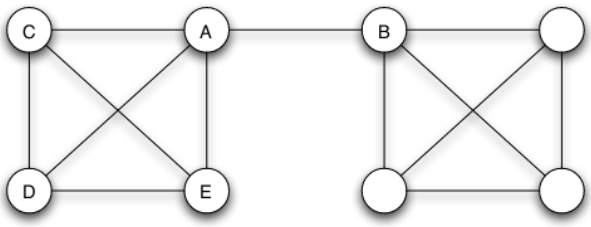



Figura 14: L'arco tra A e B rappresenta un bridge.

Questa definizione può essere estesa al caso in cui un bridge rimosso non separi completamente due nodi in componenti distinte, ma semplicemente ne allunghi la distanza di almeno 2 archi. In questo caso l'arco viene denominato **local bridge**. Diciamo che lo **span** di un local bridge è la distanza tra i due endpoints nel caso in cui l'arco venga rimosso.

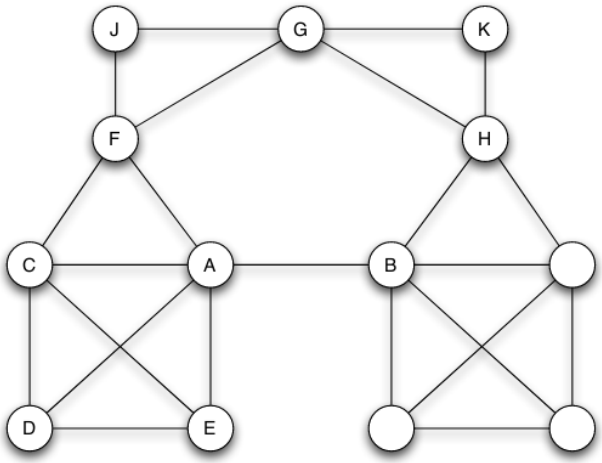



Figura 15: L'arco tra A e B rappresenta un local bridge.

Il ruolo di un local bridge, in particolare quelli con un grande span, è spesso simile a quello di un bridge normale, ovvero fornire tra due nodi un passaggio di informazioni che altrimenti non avrebbero la possibilità di ottenere; inoltre gli endpoints di un bridge o local bridge ricevono le informazioni provenienti da altre regioni prima degli altri nodi. I semplici bridges però non sono molto diffusi nelle reti reali.

### 3.2.2 The Strong triadic closure Property

Tornando all'esperimento di Granovetter, possiamo ora capire come sia importante l'idea che i collegamenti nelle social network possano avere una diversa "forza", al di là di cosa questa rappresenti. Per semplicità rappresentiamo questa proprietà in maniera binaria, definendo gli **strong ties**, che corrispondono agli amici, ed i **weak ties**, relativi ai conoscenti. Rappresentiamo questa informazione aggiungendo un'etichetta sugli archi indicando il tipo di legame (W o S).

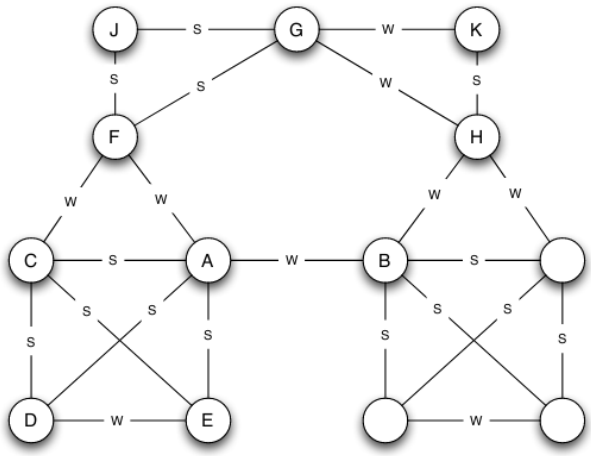



Figura 16: La rappsentazione della forza degli archi.

Rivedendo in questa ottica la proprietà di triadic closure, possiamo allora affermare che se un nodo A ha collegamenti con i nodi B e C, un arco tra B e C si formerà specialmente nel caso in cui gli altri due archi siano strong ties. Granovetter propose una versione più formale di questo concetto, definendo la **Strong triadic closure**:

> Un nodo A viola la proprietà se esso ha due strong ties con i nodi B e C, ma questi non hanno alcun collegamento tra loro.

Questa nuova proprietà è però troppo forte perché ci si aspetti che valga per tutto un grafo di grandi dimensioni, ma offre la possibilità di ragionare sulle conseguenze di strong e weak ties.

### 3.2.3 Local Bridges and Weak Ties

Possiamo ora correlare il concetto locale di strong o weak tie con quello globale di local bridge. Non c'è un collegamento diretto tra le due nozioni, ma usando la proprietà di triadic closure possiamo affermare quanto segue:

> Se il nodo nella rete soddisfa la proprietà di Strong triadic closure ed è coinvolto in almeno due strong ties, allora ogni local bridge in cui esso è coinvolto deve essere un weak tie.

Dimostriamo questa proprietà con un semplice teorema:

> *Ipotesi*: il nodo A soddisfa la Strong triadic closure ed è coinvolto ad almeno due strong ties; inoltre l'arco con il nodo B è un local bridge di tipo strong.

*Dimostrazione*: se A è coinvolto in almeno due strong ties e AB è uno di questi, deve essercene un altro con un nodo che chiamiamo C. Dato che AB è un local bridge, i due endpoint non hanno amici in comune, quindi non esiste l'arco BC. Ma ciò contraddice la Strong triadic closure, che afferma che essendo AB ed AC due strong ties, deve esistere l'arco BC. Questo contraddice l'ipotesi iniziale dell'esistenza di un local bridge di tipo strong.

Il teorema è quindi dimostrato per contraddizione e lega il concetto locale della forza dei collegamenti con quello globale dei local bridges.

Si può osservare che generalmente i local bridges sono weak ties, perché in caso contrario la triadic closure porterebbe a delle scorciatoie tra i due endpoints dell'arco.

## 3.3 Tie Strenght and Network Structure in Large-Scale Data

La connessione tra la forza dei collegamenti e le proprietà strutturali delle social networks sottostante portano a ipotesi intriganti sull'organizzazione delle social networks nella realtà. L'ipotesi di Granovetter non aveva abbastanza dati realistici per essere dimostrata e le sue assunzioni rimasero senza prove per diverso tempo. La situazione cominciò a cambiare quando divennero disponibili le prime tracce delle comunicazioni digitali, delle who-talks-to-whom network in cui si evidenzia la struttura della rete, in particolare le coppie che conversano, e la durata di tali conversazioni. Questo introduce l'idea che non si debba ragionare solamente in termini di strong e weak ties, ma che ci sia una "scala di grigi" per rappresentare la forza di un collegamento, come ad esempio il numero di minuti di conversazione tra due persone.

In uno dei maggiori studi di questo tipo, Onnela et al. fecero un esperimento della durata di 18 settimane, usando i dati di una compagnia telefonica che copriva circa il 20% della popolazione nazionale, e trovarono una giant component contenente l'84% degli utenti.

### 3.3.1 Generalizing the Notions of Weak Ties and Local Bridges

Sia per il concetto di "strength of a tie" sia per quello di "local bridge" vogliamo arrivare ad una misura che non sia binaria, ma esprima una gradazione del valore effettivo, in modo da poter avere dei risultati ragionevoli nello studio di grandi dati su una rete reale. Possiamo quindi indicare la forza di un collegamento come il totale dei minuti di conversazione tra i due endpoints

dell'arco. Inoltre poiché i local bridge sono una quantità minore di tutti gli archi presenti, si può semplificare la definizione in modo da poter considerare alcuni archi come se fossero "quasi" dei local bridges; per fare ciò definiamo il **neighborhood overlap** di un arco AB come il rapporto tra il numero di nodi vicini sia di A che di B ed il numero di nodi che sono vicini di almeno uno tra A e B (con A e B esclusi), ovvero:

$$O_{AB} = \frac{|N(A) \cap N(B)|}{|N(A) \cup N(B) \ / \{A,B\}|}$$

dove N(X) indica il numero di vicini del nodo X.

Usando questa definizione otteniamo che il valore del rapporto è zero quando quando il numeratore è zero, ovvero l'arco è un local bridge; quando il valore è prossimo a zero, possiamo invece considerare l'arco come un "quasi" local bridge.

### 3.3.2 Empirical Results on Tie Strength and Neighborhood Overlap

Usando la definizione di neighborhood overlap, possiamo farci alcune domande qualitative sulle ipotesi di Granovetter. In primis possiamo cercare la relazione tra il neighborhood overlap di un arco e la sua forza. Nella figura 17 vediamo che tra le due quantità c'è una relazione di proporzionalità diretta.

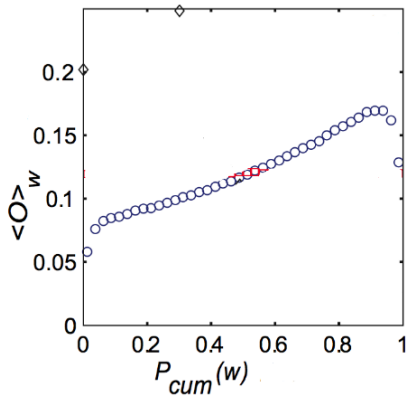



Figura 17: Un grafico del rapporto tra neighborhood overlap e tie strength.

Questa connessione però è espressa a livello locale, ma sarebbe interessante sapere cosa comporti a livelli globale. Sempre Onnela et al. cominciarono allora cancellare gli archi in ordine di strength, rimuovendo prima i più forti, e notarono che la giant component cominciava a ridursi gradualmente. Provarono poi a rimuoverli in ordine inverso e scoprirono che la giant component

si riduceva in maniera più drastica, fino a spezzarsi in seguito alla rimozione di alcuni archi particolari.

## 3.4 Tie Strength, Social Media, and Passive Engagement

Negli ultimi anni le relazioni sociali si sono sempre più spostate online, cambiando così il modo di accedere e mantenere la propria rete sociale. Ad esempio, il tenere conto di tutte le amicizie era prima qualcosa di implicito, mentre online si ha una lista ben precisa dei propri contatti. Che effetto può avere questa considerazione sulla struttura della rete?

La forza dei collegamenti può offrire una prospettiva importante su questo dilemma, rispondendo a come l'attività sociale online sia distribuita nei collegamenti e nella loro forza.

### 3.4.1 Tie Strength on Facebook

Per rispondere a questa domanda i ricercatori hanno cominciato a studiare diversi siti di social media. Ad esempio in Facebook, Cameron Marlow et al. analizzarono i collegamenti di amicizia in ogni profilo e si chiesero fino a che punto questo collegamento fosse effettivamente usato per una interazione sociale. In altre parole cercarono la presenza di strong ties tra amici, e per farlo divisero i collegamenti in tre categorie principali, trovate dopo un mese di osservazione dei dati:

1. un collegamento rappresenta una *comunicazione reciproca* se entrambi si scambiano messaggi;

2. un collegamento rappresenta una *comunicazione a senso unico* se un utente ha mandato uno o più messaggi all'amico (sia se questi messaggi fossero reciproci o meno);

3. un collegamento rappresenta una *comunicazione mantenuta* se un utente ha seguito le informazioni dell'altro, come ad esempio cliccare su un link condiviso, con o senza scambio di messaggi.

Si noti che queste categorie non sono mutualmente esclusive. L'immagine 18 rappresenta i risultati di questo studio.

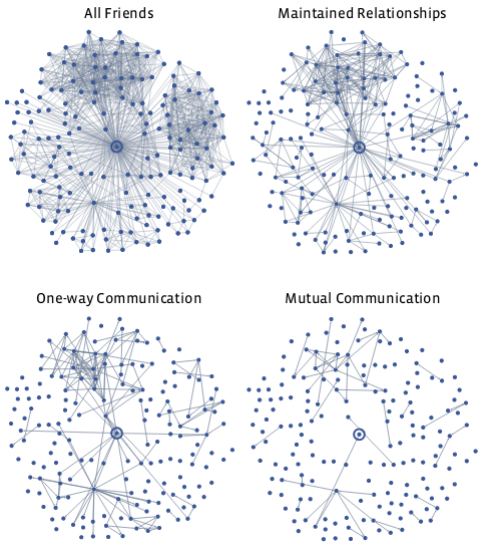



Figura 18: Quattro rappresentazioni della rete dei vicini di un utente Facebook.

Inoltre l'immagine 19 si può evincere che chi possiede una grande quantità di amicizie (più di 500) in realtà ha una comunicazione reciproca con un numero di persone compreso tra 10 e 20 e ne segue meno di 50. Per questo Marlow ed i suoi colleghi arrivarono alla conclusione che i social media portino ad una sorta di *impegno passivo*, nel quale un rapporto è mantenuto leggendo notizie dell'amico, senza comunicare con lui.

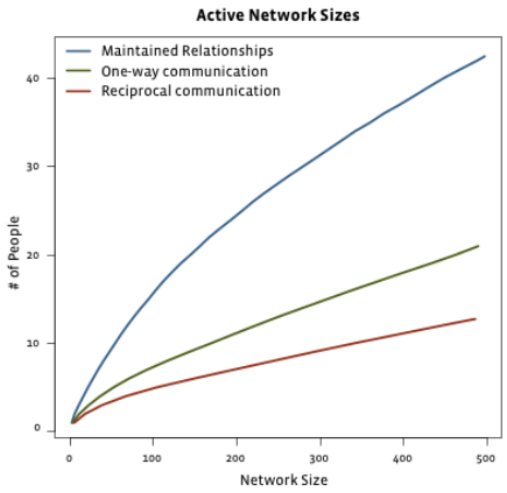



Figura 19: Il rapporto tra il tipo di relazione e la dimensione della rete degli utenti Facebook.

### 3.4.2 Tie Strength on Twitter

Degli studi simili sono stati effettuati anche sulla piattaforma di Twitter, dove è possibile seguire altri utenti anche senza il loro diretto consenso. Questo tipo di relazione è stata considerata come un weak tie, mentre lo scambio di messaggi verso altri utenti rappresenta uno strong tie. Su queste basi, Huberman, Romero e Wu analizzarono i dati da loro raccolti e scoprirono che, come nel caso di Facebook, anche per gli utenti con molti weak ties, il numero di strong ties rimane decisamente inferiore, in questo caso stabilizzandosi a 50, su 1000 utenti seguiti, come evidenziato in figura 20.

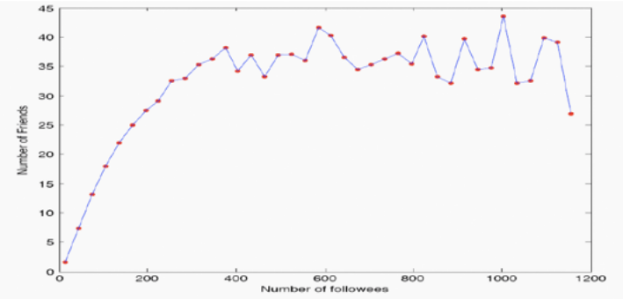



Figura 20: Il rapporto tra il numero di strong ties ed il numero di followees su Twitter.

Una delle spiegazioni possibili per questo fenomeno è il fatto che per mantenere attive delle relazioni serve tempo ed energia, e per questo raggiunto un certo limite la quantità di strong ties si arresta; al contrario, la formazione di weak ties richiede molte meno risorse e per questo il numero può crescere maggiormente.

## 3.5 Closure, Structural Holes, and Social Capital

Fino ad ora la discussione ha portato ad una visione della rete sociale come formata da diversi gruppi molto legati dai weak ties. Abbiamo parlato dell'importanza di quest'ultimi, ma bisognerebbe anche discutere del ruolo che possono avere i nodi all'interno di un grafo. All'interno di una rete, l'accesso agli archi che dividono i gruppi non è equamente distribuito tra tutti i nodi e questo porta a diversità nella loro esperienza nella struttura.

Ma qual è l'effetto di questa eterogeneità? Seguendo gli studi di Ron Burt possiamo trovare le differenze tra i nodi che stanno al centro di un gruppo molto connesso ed i nodi che invece si interfacciano a diversi gruppi.

### 3.5.1 Embeddedness

Consideriamo il nodo A, i cui vicini sono soggetti ad una considerabile triadic closure e per questo A ha un alto clustering coefficient. Per parlare della struttura attorno ad A è necessario introdurre un nuovo concetto legato agli archi, chiamato **embeddedness** e definito come il numero di vicini comuni ai suoi due endpoints. Per precisazione, questa definizione corrisponde al numeratore della definizione di neighborhood overlap. In figura 21 vediamo che gli archi di A hanno una significante embeddedness. Uno studio ha dimostrato che se due individui sono connessi da un arco embedded è più facile per loro avere fiducia reciproca. Inoltre la presenza di una mutua amicizia mette in vista i due individui agli occhi degli amici comuni. Lo stesso Granovetter afferma che la sua mortificazione nel tradire un caro amico sarebbe molta anche nel caso non venisse scoperto, crescerebbe nel caso l'amico lo scoprisse e sarebbe insopportabile se un amico comune lo sapesse e lo riferisse ad altri.

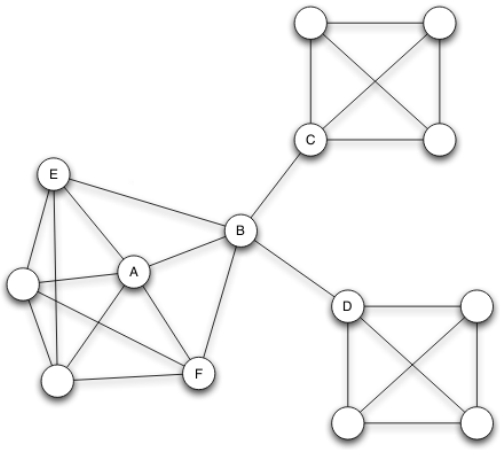



Figura 21: Il contrasto tra gruppi densamente legati ed i collegamenti che definiscono i confini mostrato nelle diverse posizioni dei nodi A e B.

Questa preoccupazione non esiste nel caso di archi con zero embeddedness, poiché nessuno sa delle relazioni che avvengono. Per questo le interazioni di B con C e D sono più rischiose di quella con A, perché potenzialmente esposte a norme contraddittorie e diverse a seconda del gruppo con cui il nodo si interfaccia.

### 3.5.2 Structural Holes

Finora abbiamo citato gli svantaggi del nodo B per la sua posizione, ma uno studio caratterizzato dal lavoro di Burt ha mostrato che vi sono altrettanti

vantaggi. Il tipico ambiente per questa argomentazione è la rete sociale di una azienda, composta da diversi gruppi di persone che cercano di collaborare per un obiettivo comune. Immaginando la figura 21 come una rete delle collaborazioni tra i diversi manager dell'azienda, possiamo affermare che il nodo B, con i suoi molti local bridges, rappresenta uno **structural hole**, ovvero uno spazio vuoto nella rete tra due gruppi che altrimenti non avrebbero modo di comunicare.

Questa posizione offre a B diversi vantaggi, fra cui l'avere accesso per primo alle informazioni scambiate tra i gruppi, il poter usare queste diverse informazioni provenienti da diversi gruppi per costruire concetti sintetizzati da diverse idee, oppure ancora lo svolgere la funzione di "gatekeeping", controllando come C e D hanno accesso al suo gruppo e filtrando le informazioni di C e D per il suo gruppo. Questo può essere un bene per B, ma non per l'organizzazione di cui fa parte.

In conclusione quindi la posizione del nodo B lo lega di meno ad un singolo gruppo e fornisce meno protezione dalla presenza di amici comuni, gli fornisce accesso ad informazioni che risiedono in diversi gruppi con l'opportunità di regolarne il flusso e di sintetizzarne altre.

### 3.5.3 Close and Bridging as Forms of Social Capital

Tutti questi concetti sono presentati con l'idea di un individuo o gruppo che trae vantaggio dalla struttura sottostante della struttura sociale; per questo motivo essi sono legati al concetto di **social capital**, un termine molto usato ma difficile da definire.

La sua formulazione può suggerire il suo ruolo come una parte di un array di diverse forme di capitale, ognuna delle quali serve una risorsa tangibile o meno che può essere mobilitata per eseguire un certo compito. Il termine è spesso associato a quelli di *physical capital*, cioè le tecnologie che permettono di svolgere un certo compito, *human capital*, ovvero le abilità degli individui, *economic capital*, ossia le risorse monetarie, e *cultural capital*, cioè le risorse accumulate da una cultura che vanno oltre il singolo individuo. Spesso però vengono osservate due principali accezioni con cui si fa uso del termine social capital: nella prima esso è spesso visto come la proprietà di un gruppo, che distingue i gruppi che funzionano meglio da quelli che rendono meno, oppure è considerato come proprietà di un individuo, che varia a seconda della sua posizione nella struttura sociale sottostante. Nella seconda accezione invece è basata sul riferirsi ad esso come una proprietà intrinseca di un gruppo, oppure derivante dalle connessioni esterne del gruppo stesso.

Una visione così generica non permette di definire quale sia la struttura sociale migliore per incrementare il social capital, ed inoltre causa la nascita

di diversi punti di vista sul tema. Ad esempio Coleman afferma che il social capital enfatizzi i benefici della triadic closure e dell'embeddedness, mentre invece Burt sostiene che esso causi una tensione tra closure, riferendosi al concetto di Coleman, e brokerage, ossia l'abilità di fungere da intermediario tra gruppi diversi.

## 3.6 Advanced Material: Betweenness Measures and Graph Partitioning

Fino ad ora abbiamo parlato dei gruppi di nodi strettamente legati in una rete, ma senza fornirne una definizione precisa, poiché essa può dipendere dallo scenario analizzato. Ci sono però casi in cui una definizione precisa torna utile, specialmente nel caso in cui ci affacciamo ad un dataset di una rete e vogliamo identificare i gruppi densamente connessi.

Cerchiamo quindi un metodo che permetta di prendere una rete e dividerla in insiemi molto connessi, con interconnessioni più sparse tra le diverse regioni. Ci riferiamo a questo problema usando il nome di **graph partitioning** o di **community detection**, ed alle parti in cui la rete viene divisa con **regions**.

Per dare un senso a questo discorso possiamo fare due esempi: nella figura 23 vediamo una collaboration network tra diversi autori scientifici, con diversi gruppi molto connessi ed alcuni individui ai confini tra i gruppi; nella figura 23 vediamo invece il club di karate studiato da Zachary, in cui una disputa tra il presidente (nodo 34) ed l'istruttore (nodo 1) porta alla separazione della rete in due regioni principali.

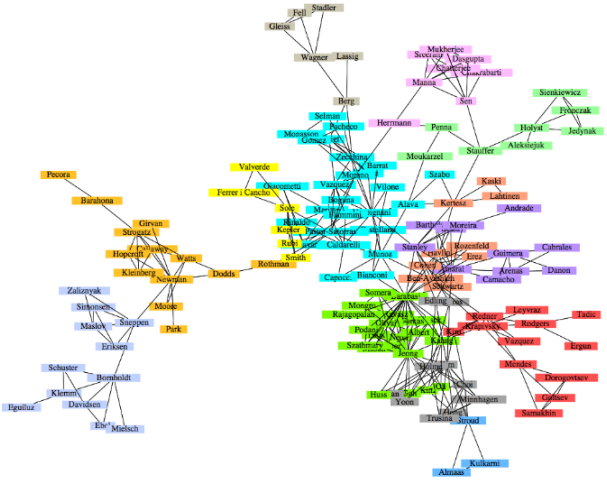


Figura 22: Una rete dei coautori di fisica e matematica applicata.

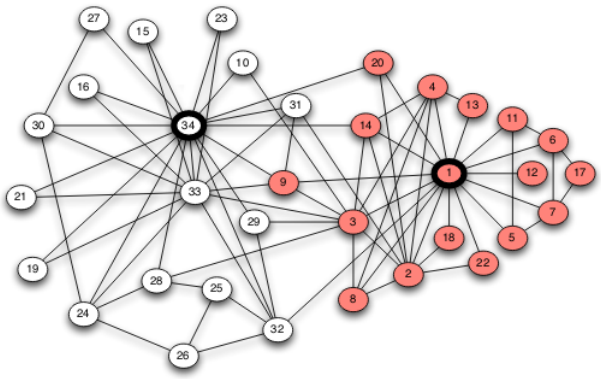



Figura 23: Il karate club studiato da Zachary.

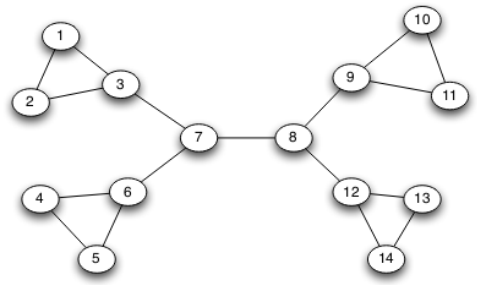
### 3.6.1 A method for Graph Partitioning

Per dividere il grafo sono state proposte diverse tecniche ed è importante analizzarne per capirne i diversi stili che motivano la loro ideazione.

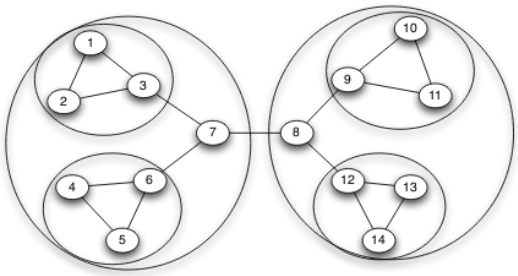
**General Approaches to Graph Partitioning.** Una classe di metodi si concentra sul rimuovere i collegamenti tra le regioni densamente connesse; una volta rimossi i collegamenti, la rete comincia a dividersi in grandi parti, nelle quali viene ripetuto il processo di rimozione ed il processo continua. Questo tipo di metodi viene chiamato **divisive** ed è di tipo top-down.

Un altro metodo è quello di concentrarsi sulle parti più densamente connesse della rete, invece sulle connessioni ai loro confini, trovando i nodi che fanno parte della stessa regione ed unendoli insieme. A questo punto la rete risulta formata da diversi blocchi di nodi uniti, e si procede cercando altri blocchi da unire tra loro. Questa tipologia viene chiamata **agglomerative** ed è un approccio di tipo bottom-up.

Nella figura 24 possiamo vedere come questo procedimento porti al formarsi di regioni annidate tra loro. I metodi divisivi inizieranno a suddividere il grafo nelle regioni più grandi, per poi trovare all'interno di esse quelle minori; i metodi aggregativi invece seguiranno il processo contrario.




(a) *A sample network*




(b) *Tightly-knit regions and their nested structure*

Figura 24: Molte reti mostrano apparenti regioni molto connesse e possono avere gruppi annidati.

**The Notion of Betweenness.** Tra i diversi metodi ideati è importante citare quello di tipo divisivo proposto da Girvan e Newman, ma per farlo serve prima introdurre il concetto di **betweenness**. In genere per dividere due regioni la prima cosa che viene in mente di fare è cercare gli archi bridge oppure local bridge, poiché questi spesso dividono le diverse regioni della rete. Ma in presenza di più bridges, quale bisogna rimuovere per primo? E se invece nel grafo non ci fossero local bridges perché ogni arco appartiene ad un triangolo, ma ci fosse comunque una divisione naturale in regioni?

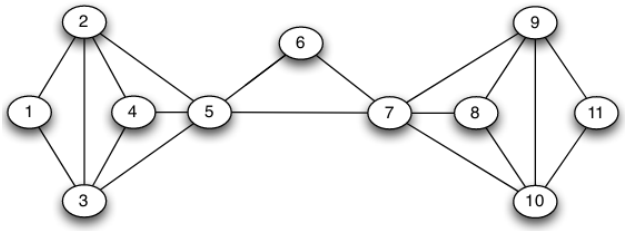



Figura 25: Un grafo può avere regioni molto connesse divise tra loro anche quando non sono presenti bridges o local bridges a dividerle.

I local bridges sono importanti perché fanno parte del cammino minimo tra coppie nodi situati in parti diverse della rete. Possiamo quindi definire la nozione astratta di **traffic** sulla rete e vedere quali archi ne trasportano di più. Il traffico si può definire nel sequente modo:

> Per ogni coppia di nodi A e B connessi da un cammino, immaginiamo di avere una unità di fluido "flow" tra gli archi da A a B. Il flow tra A e B si divide equamente tra tutti i cammini *minimi* tra i due nodi, quindi se esistono $k$ cammini minimi, $1/k$ di flow passa per ogni cammino.

Definiamo la betweenness di un arco come il numero totale di flow che traporta, considerando tutte le coppie di nodi che usano quell'arco.

$$B(i) = \sum_{j<k} \frac{d_{jk}(i)}{d_{jk}}$$

Questa definizione è spesso usata dai sociologi in riferimento ai nodi anziché agli archi, ma il concetto di base rimane lo stesso.

**The Girvan-Newman Method: Successively Deleting Edges of High Betweenness.** Basandoci quindi sul concetto di betweenness possiamo osservare che gli archi più importanti per connettere diverse regioni di un grafo sono quelli con un flow maggiore. Questo approccio porta a definire il **metodo di Girvan-Newman** come segue:

1. trovare gli archi con maggiore betweenness e rimuoverli dal grafo. Il processo può portare il grafo a dividersi in diverse componenti; se ciò succede, questo è il primo livello delle regioni nel partizionare il grafo;

2. ricalcolare la betweenness e nuovamente rimuovere gli archi con il valore maggiore. Questa procedura può dividere il grafo in altre componenti, che sono annidate nelle regioni trovate precedentemente;

3. ripetere i passi precedenti finché non rimangono archi nel grafo.

Nella presentazione del loro lavoro Girvan e Newman usarono alcuni datasets reali per mostrare l'algoritmo, tra cui il karate club di Zachary. In questo preciso esempio però un nodo (il numero 9 per precisione) trovò la collocazione sbagliata, venendo legato al presidente invece che all'istruttore come avvenne nella realtà. L'approccio che invece aveva presentato Zachary stesso era basato sul trovare il **minimum cut** nel grafo, ovvero cancellare gli archi di forza minima totale a separare due nodi, ma il risultato fu lo stesso dell'algoritmo Girvan-Newman.

### 3.6.2 Computing Betweenness Values

Per eseguire il metodo di Girvan-Newman occorre poter calcolare in ogni step il valore più alto di betweenness tra tutti gli archi. Ma per fare questo è necessario poter analizzare l'insieme di *tutti* i cammini minimi tra le coppie di nodi, ed in presenza di una vasta rete ciò può risultare lento e complesso. Un modo efficiente per calcolare la betweenness è quello di applicare l'algoritmo di breadth-first search; dopodiché con un approccio top-down si calcolano i cammini minimi; poi in maniera bottom-up si calcola come il flow sia distribuito verso gli altri nodi. Infine semplicemente si sommano i risultati di tutti i nodi per ottenere la betwenness di ogni arco.

# 4 Networks in Their Surrounding Contexts

Fino ad ora abbiamo visto le dinamiche interne ad una rete, ma è importante sapere che anche il **contesto circostante**, ovvero tutti i fattori che esistono al di fuori dei nodi e degli archi di una rete, ha effetti su come la struttura delle rete si evolve.

## 4.1 Homophily

Una delle nozioni di base che regolano la struttura delle reti sociali è quella di **homophily**, il principio per cui si tende ad essere simili ai propri amici. In genere i propri amici non sembrano dei campioni casuali della popolazione, ma hanno invece caratteristiche simili per quanto riguarda razza ed etnia, età, posto dove vivono, lavoro, interessi ed opinioni. Chiaramente esistono amicizie che non rispettano questi canoni, ma nel complesso i collegamenti in una rete sociale tendono ad aggregare persone con caratteristiche simili.

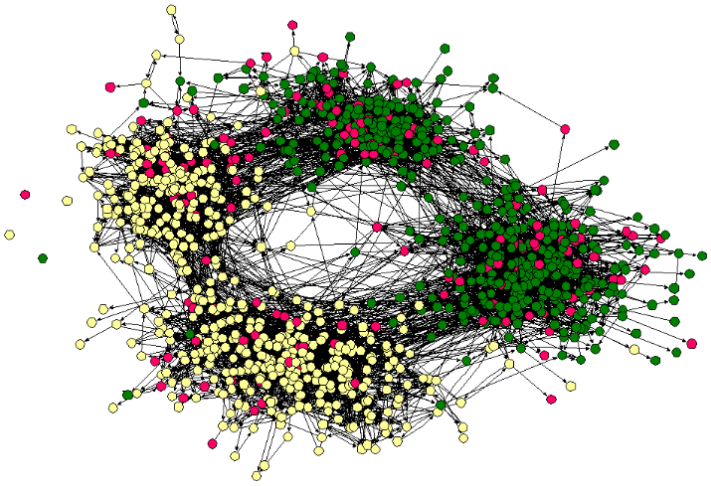


Figura 26: Un esempio di homophily tratto da una scuola media e superiore. Si possono notare due divisioni principali: la prima per razza (mostrata dai colori dei pallini) ed una per amicizia nella scuola media e superiore.

L'homophily ci offre una prima visione su come il contesto esterno di rete può guidare le informazioni dei suoi collegamenti. Ad esempio, quando un'amicizia nasce perché si è presentati da un amico comune il collegamento ha origini intrinseche alla rete, mentre se un'amicizia nasce dal frequentare la stessa scuola o lavoro ha senso guardare il contesto in cui il nodo è situato.

Ovviamente ci sono interazioni intrinseche e relative al contesto per ogni singolo collegamento, ed esse operano in contemporanea nella rete. Ad esempio per motivare la triadic closure abbiamo parlato di opportunità ed incentivo, ma possiamo ora aggiungere che un amico di un nostro amico probabilmente ha caratteristiche simili alle nostre per homophily, e questo potrebbe portare alla chiusura del triangolo.

### 4.1.1 Measuring Homophily

Quando vediamo una rete come quella della figura 26 è importante chiedersi se una divisione del genere sia effettivamente presente o se emerga solamente dal tipo di rappresentazione.

Può essere difficile ragionare su una misura partendo da una rete grande, per questo usiamo la rete in figura 27, in cui i nodi sono divisi in due soli tipi, riconducibili ad esempio ai due sessi. Partiamo chiedendoci cosa vorrebbe dire per la rete non mostrare homophily per lo stesso genere? Significherebbe che la percentuale di amicizie di un genere sarebbe proporzionale alla distribuzione totale della rete per ogni individuo, ovvero che se assegnassimo ad ogni nodo un genere con una probabilità relativa alla distribuzione effettiva nella rete, il numero di collegamenti tra generi diversi non cambierebbe significativamente.

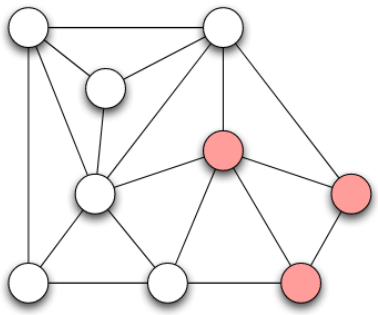


Figura 27: Una piccola rete per studiare la misura dell'homophily.

Supponiamo quindi di avere una rete in cui una frazione $p$ di tutti gli individui è maschile ed una frazione $q$ è invece femminile. Dato un arco in questa rete, se assegnassimo ad ogni nodo la proprietà maschile con probabilità p e quella femminile con probabilità q, la probabilità che gli endpoints dell'arco siano entrambi maschi è $p^2$, che siano entrambi femmine è $q^2$, che siano uno maschio ed uno femmina è $2pq$.

Con questi presupposti possiamo allora formalizzare il test per l'homophily come segue:

> Se la frazione di archi tra generi diversi è significativamente minore di $2pq$ allora vi è presenza di homophily.

Nella figura 27 ad esempio 5 dei 18 archi sono tra generi diversi, $p$ vale $2/3$ e $q$ vale $1/3$, troviamo che $2pq$ vale $8/18$. Ci si aspetterebbe quindi di vedere 8 archi tra generi diversi invece che solamente 5, per questo l'esempio mostra la presenza di homophily.

Bisogna fare però alcune considerazioni finali importanti:

- è importante definire precisamente il termine "significativamente minore" in base al contesto di studio;
- può succedere che una rete mostri un numero di archi tra diversi tipi significativamente maggiore di $2pq$. In questo caso si parla di **inverse homophily**;
- quando ci sono più di due caratteristiche da analizzare, effettuiamo una versione generalizzata del calcolo, dicendo che un arco è **heterogeneous** se connette due nodi che sono diversi secondo le caratteristiche presenti. Ci si chiede poi se il numero di archi heterogeneous è comparabile con cosa otterremmo assegnando le caratteristiche in maniera casuale ai nodi della rete, usando le proporzioni dei dati reali come probabilità.

## 4.2 Mechanisms Underlying Homophily: Selection and Social Influence

Il fatto che un individuo tenda a formare collegamenti con altri simili fornisce una informazione strutturale sulla rete, ma non offre spiega il meccanismo sottostante che porta al formarsi di questi legami.

Nei casi di caratteristiche immutabili come razza o etnia, la tendenza a formare legami con altri simili è chiamata **selection**. La selezione può agire in diverse scale e con diversi livelli di intenzionalità. Ad esempio lo scegliere le proprie amicizie all'interno di un gruppo ristretto comporta una scelta attiva; a livello più globale la selezione può essere più implicita, come nel caso di persone che vivono nello stesso vicinato, frequentano la stessa scuola o lavorano per la stessa azienda.

Le caratteristiche prima citate sono immutabili ed hanno un ruolo importante nelle connessioni che un individuo forma durante la sua vita. Gli attributi variabili come comportamenti, attività, credenze, interessi ed opinioni hanno invece un ruolo molto più complesso nel formare legami: le persone possono mutare questi attributi per avvicinarsi al comportamento dei loro amici. Questo processo è descritto come **socialization** e **social influence**.

### 4.2.1 The Interplay of Selection and Social Influence

Quando si analizzano le caratteristiche mutabili in una rete, può essere difficile capire quanto intervenga la selezione e quanto l'influenza sociale in un individuo. Per questo si effettuano degli studi **longitudinal**, in cui le interazioni ed evoluzioni sono osservate e studiate in un periodo di tempo. Ciò permette di capire quanto influisca la selezione e quanto l'influenza in una rete, e di agire in conseguenza per "manipolare" la diffusione.

Per fare un esempio, studiando una rete di studenti si nota che alcuni gruppi fanno uso di droghe. Se l'homophily in questo ambito nasce dall'influenza sociale, si può creare un programma di recupero che diffondendosi può ridurre il problema; se invece l'homophily deriva dalla selezione, si osserverà che il programma porterà l'individuo a cambiare le proprie amicizie verso gruppi diversi, ma senza influenzare le scelte degli altri.

Un altro esempio può essere lo studio di Christakis e Fowler sulla relazione tra l'obesità e una social network, notando che l'homophily influisce parecchio. Ma il problema è capire quali tra diverse ipotesi spiega questo fatto:

1. è per effetto della selezione?
2. è perché in realtà l'homophily lavora su altre caratteristiche che sono correlate all'obesità?
3. è perché il cambiamento dello stato di obesità di un amico ha influenzato lo stato di obesità attuale o futuro dell'individuo stesso?

Dagli studi effettuati emerse che nonostante i motivi principali siano i primi due punti, vi è anche una influenza forte della terza motivazione: l'ipotesi è che una condizione con un forte aspetto comportamentale tenda a **contagiare** gli altri.

Questi esempi mostrano che gli studi sull'homophily non sono fini a se stessi, ma spesso sono il punto di partenza per porsi delle domande più profonde sui meccanismi alla base della rete.

## 4.3 Affiliation

Abbiamo visto come alcuni fattori esterni possano influenzare le attività dei nodi all'interno della rete, ma in realtà è possibile includere il contesto esterno nella rete stessa, trasformandola in una rete più grande composta da nodi che rappresentano gli individui ed il contesto.

Si potrebbe rappresentare qualsiasi contesto in questo modo, ma ci occuperemo maggiormente delle **attività** a cui una persona prende parte, come

essere compreso in una scuola, un lavoro od una organizzazione, avere determinati hobby o ancora frequentare determinati posti. Il termine usato per riferirsi a queste attività è **foci**, ovvero i punti focali di interazione sociale.

### 4.3.1 Affiliation Networks

Possiamo rappresentare la partecipazione di un gruppo di persone ad un insieme di foci come un semplice grafo. L'individuo A partecipa al focus X se vi è un arco che collega i due nodi. Questo tipo di grafo è chiamato **affiliation network**; più in generale possiamo dire che esso è un grafo bipartito in cui i nodi sono divisi in due tipi diversi ed in cui gli archi non collegano nodi dello stesso tipo. Vediamo un esempio nella figura 28.

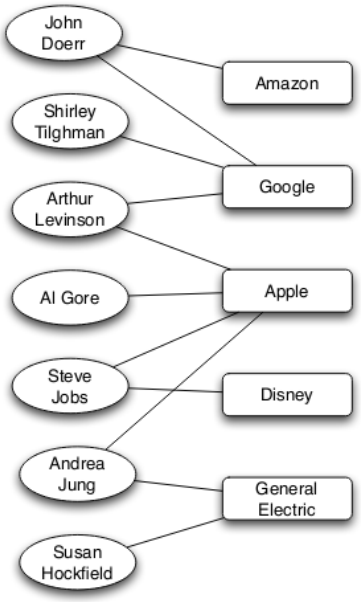



Figura 28: Un esempio di affiliation network.

Le affiliation networks sono studiate quando i ricercatori vogliono capire i pattern della partecipazione in determinate attività.

### 4.3.2 Coevolution of Social and Affiliation Networks

Chiaramente sia le social networks sia le affiliation networks cambiano nel tempo: nuove amicizie si formano e le persone si interessano a nuove foci. Questi cambiamenti rappresentano una **coevolution** che mostra il legame tra selection e social influence: se due persone partecipano allo stesso focus,

hanno la possibilità di diventare amici; se due individui sono amici, sono portati ad interessarsi allo stesso focus.

Per rappresentare entrambe le nozioni si usano le **social-affiliation networks**, dei grafi formati da due tipi di nodi che rappresentano gli individui ed i foci, e da due tipi di archi: un collegamento tra due individui rappresenta un'amicizia, mentre un collegamento tra un individuo ed un focus rappresenta un'attività.

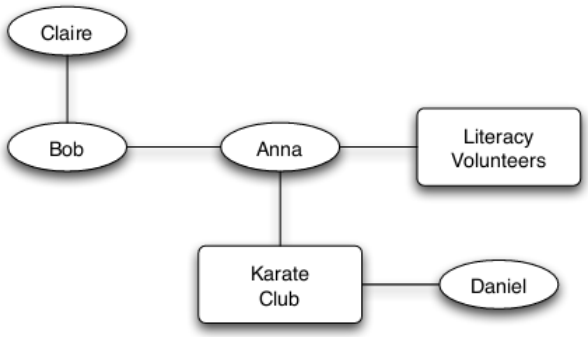


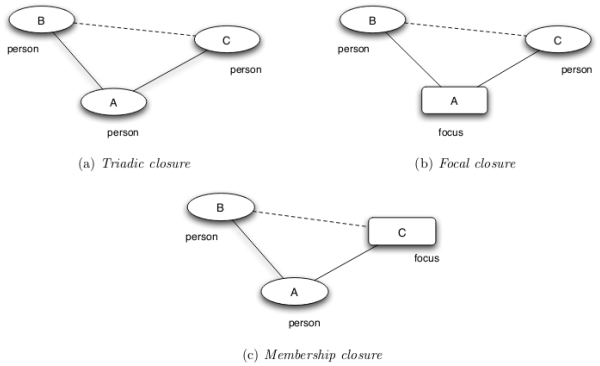

Figura 29: Un esempio di social-affiliation network.

Una volta capito come rappresentare questo tipo di rete, possiamo interessarci ai diversi meccanismi che portano alla formazione dei collegamenti, vedendoli come un tipo di **closure process**. In particolare, supponiamo di avere due nodi B e C con un vicino A in comune e che si formi un arco tra B e C. Osservando la figura 30 vediamo come si siano tre possibili interpretazioni di questa situazione, a seconda che i nodi siano individui o foci:

1. se A, B e C sono individui, la formazione di un arco tra B e C è riconducile alla **triadic closure**;

2. se B e C sono individui ed A è un focus, la formazione di un arco tra B e C rappresenta la tendenza di creare un collegamento con individui che hanno un focus in comune. Ciò è riconducibile al principio della *selezione* e per enfatizzare questo aspetto chiameremo questo processo **focal closure**;

3. se A e B sono individui e C è un focus, allora abbiamo la formazione di una nuova affiliation: B partecipa ad un focus in cui l'amico A è già presente, in un caso di *social influence*. Chiameremo questo processo **membership closure**.




(a) *Triadic closure* (b) *Focal closure*

(c) *Membership closure*

Figura 30: Un esempio di (a) triadic closure, (b) focal closure e (c) membership closure.

Ciò che questi tre aspetti hanno in comune è la creazione di un arco tra due nodi che hanno già un collegamento tra loro.

## 4.4 Tracking Link Formation in Online Data

Abbiamo visto i principali meccanismi che portano alla formazione di nuovi collegamenti nelle reti sociali, ma essi sono difficili da quantificare con precisione. Una strategia naturale è quella di studiarli mentre operano in grandi popolazioni, dove un insieme di fattori minori può portare a risultati più grandi e complessi, e più facilmente osservabili. Tuttavia questo processo è difficilmente attuabile nella vita quotidiana, in quanto è difficile trovare un gruppo di persone e di relative foci abbastanza grande da essere osservato e studiato con cura; per questo ci si basa ancora una volta sui grandi dati resi disponibili online, con il problema però di non poter sempre estrapolare da collegamento digitale un effettivo comportamento umano.

### 4.4.1 Triadic Closure

Partendo da questi concetti, ci chiediamo quanto sia probabile che si fermi un collegamento tra due persone che hanno un amico in comune. Possiamo inoltre estendere questo quesito chiedendoci quale sia la probabilità se le due persone hanno più amici in comune, sapendo che avere più amici in comune comporta maggiore opportunità e fiducia ed una homophily più forte.

Possiamo rispondere a questi quesiti usando empiricamente i dati della rete nel seguente modo:

1. fare due snapshots della rete in due istanti diversi;

2. per ogni $k$, trovare le coppie di nodi con esattamente $k$ amici in comune nel primo snapshot, ma che non sono direttamente collegati da un arco;

3. definire $T(k)$ come la frazione di coppie che hanno formato un collegamento nel secondo snapshot. Questa è una stima della probabilità che un collegamento si formi tra individui con $k$ amici in comune;

4. rappresentare $T(k)$ come una funzione di $k$ per illustrare gli effetti degli amici in comune nella creazione di un collegamento.

Kossinets e Watts fecero uno studio di questo tipo usando un grafo who-talks-to-whom basato sullo scambio di email degli studenti di un college in un periodo di 60 giorni. Per calcolare i diversi $T(k)$ hanno comparato diverse coppie di snapshot e calcolato la media dei valori ottenuti. Il confronto tra $T(0)$ ed i valori di vari $T(k)$ risponde alle precedenti domande sulla triadic closure.

Per interpretare meglio il modello è utile confrontare i risultati dello studio con quelli di un modello semplificato, in cui gli amici in comune forniscono probabilità indipendenti di formazione di un collegamento. Supponiamo che, per una bassa probabilità $p$, ogni amico in comune tra due individui dia loro ogni giorno una probabilità indipendente $p$ di formare un collegamento. Se due persone hanno $k$ amici in comune, la probabilità che non formino un collegamento ogni giorno è (1 - $p$)$^k$. In conseguenza la probabilità che si formi un collegamento è

$$T_{baseline}(k) = 1 - (1 - p)^k$$

Nella figura 31 la linea nera mostra i risultati dello studio su dati reali, la linea tratteggiata superiore mostra i dati ottenuti dal modelli sopra citato, la linea tratteggiata inferiore mostra i dati dal modello ottenuti dalla curva 1 - (1 - p)$^k$, che trasla i risultati precedenti di una unità verso destra, dal momento che l'apporto dei primi amici in comune nei dati è ridotto.

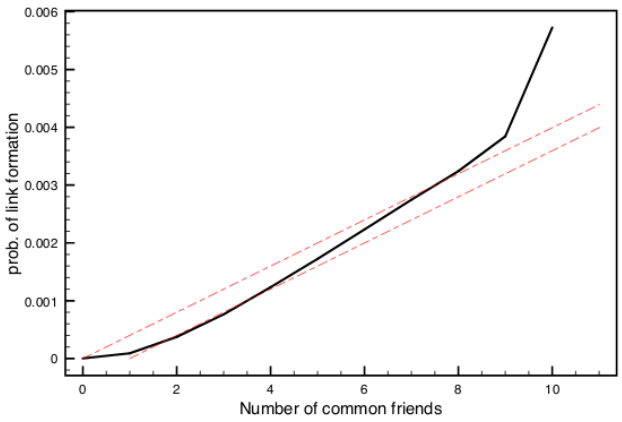



Figura 31: I risultati dello studio di Kossinets e Watts sulla triadic closure.

### 4.4.2 Focal and Membership Closure

Usando lo stesso approccio possiamo calcolare le probabilità per questi due tipi di closure. In particolare:

- per la focal closure, qual è la probabilità che due individui formino un collegamento legata al numero di foci in comune?
- per la membership closure, qual è la probabilità che una persona si interessi ad un certo focus, dato il numero di amicizie che ne fanno parte?

Per la focal closure, Kossinets e Watts aggiunsero al dataset delle email le informazioni sugli orari delle lezioni per ogni mese. Ogni corso in questo modo rappresenta un focus e due studenti sono collegati se frequentano lo stesso corso. La figura 32 mostra i dati ottenuti dallo studio sulla focal closure, seguendo lo stesso procedimento mostrato per la triadic closure.

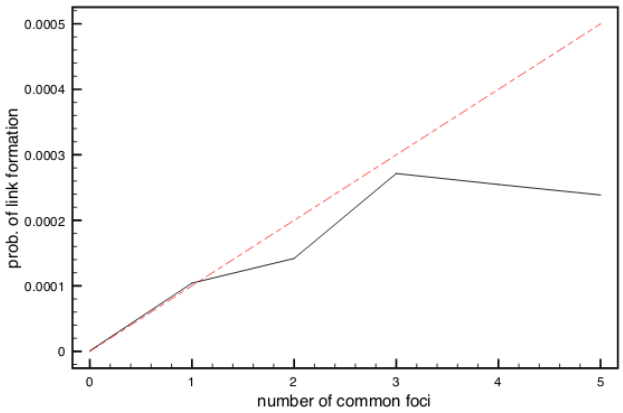



Figura 32: I risultati dello studio di Kossinets e Watts sulla focal closure.

Per quanto riguarda invece la membership closure, i dati sono stati raccolti studiando diversi domini online che mostrino interazioni tra persona e persona, e tra persona e focus. Due esempi sono la relazione tra l'iscriversi ad una comunità di LiveJournal a seconda del numero di amici già iscritti (figura 33) ed il modificare una pagina di Wikipedia a seconda del numero di amici che lo hanno già fatto (figura 34).

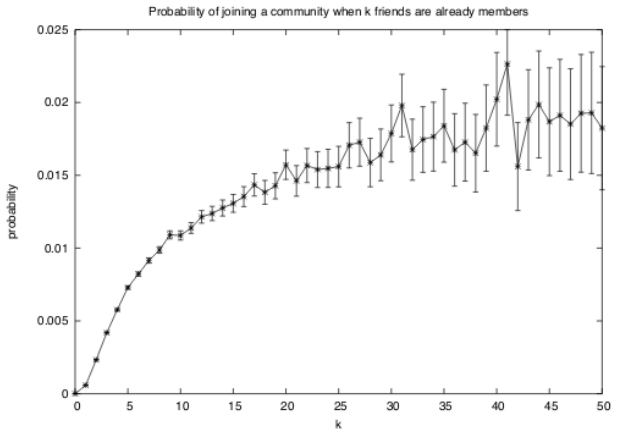



Figura 33: I risultati dello studio di Kossinets e Watts sulla membership closure sulle comunità di LiveJournal.

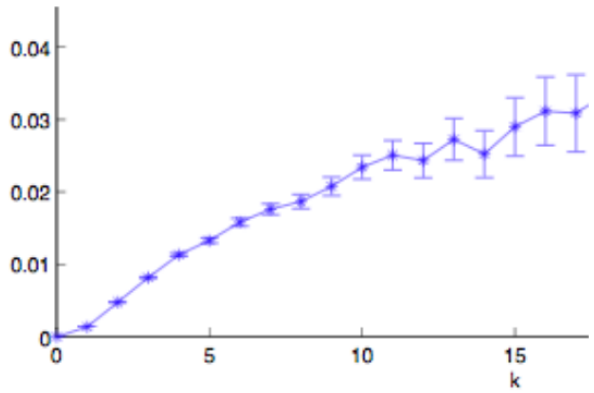



Figura 34: I risultati dello studio di Kossinets e Watts sulla membership closure sulle pagine di Wikipedia.

Ripetiamo ancora una volta che ovviamente molti diversi effetti possono operare contemporaneamente e portare a questi risultati, e tenere traccia e quantificare tutti questi aspetti è molto difficile. L'homophily stessa suggerisce che gli amici hanno diverse caratteristiche in comune, tra cui ci può anche essere l'interesse verso gli stessi foci.

### 4.4.3 Quantifying the Interplay Between Selection and Social Influence

Torniamo ora al contributo di selezione e influenza sociale nel produrre homophily, ed usando i dati di Wikipedia prima rappresentati ci chiediamo come le similitudini di comportamento tra due editori di pagine si rapportano ai pattern di interazione sociale nel tempo.

Dobbiamo quindi definire sia la rete sociale, formata dagli editori che mantengono le diverse talk pages, con un collegamento se hanno scritto sulla talk page dell'altro membro. La similarità di comportamento tra due individui A e B è invece misurata con il seguente rapporto:

$$\frac{numero\ di\ pagine\ modificate\ da\ entrambi\ A\ e\ B}{numero\ di\ pagine\ modificate\ da\ almeno\ uno\ tra\ A\ e\ B}$$

Si noti come questa definizione sia praticamente la definizione del neighborhood overlap nel grafo bipartito con persone e foci.

Ci chiediamo quindi se l'homophily nasca perché gli editor formano connessioni con chi modifica le stesse pagine (selezione), oppure è perché gli editor sono condotti agli articoli di coloro a cui parlano (influenza sociale)? Dal momento che ogni azione su Wikipedia è registrata con un timestamp, possiamo osservare i dati aumentando il tempo di un tick per ogni azione e considerando come tempo 0 il momento della prima interazione tra essi.

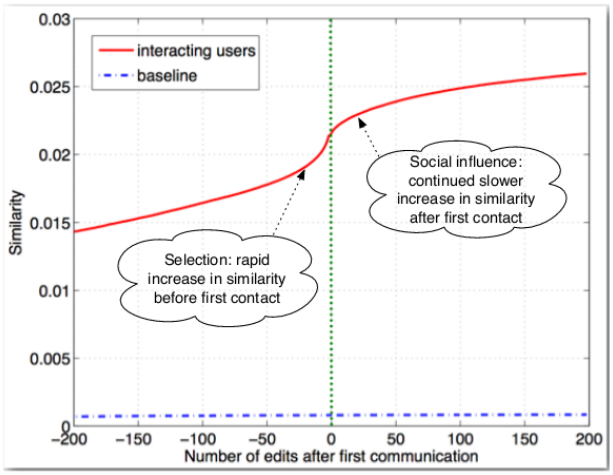


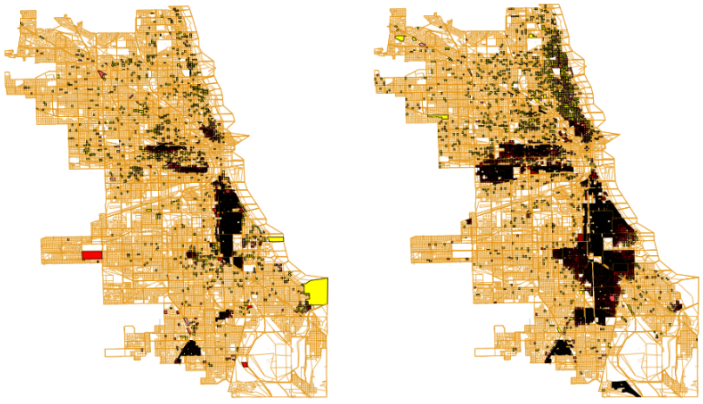

Figura 35: La similitudine media di due editori di Wikipedia.

Il risultato dello studio si può osservare nella figura 35, dove vediamo che la similarità incrementa dal momento stesso della prima interazione, a mostra come selezione ed influenza sociale cooperino. La curva non è però simmetrica relativamente allo zero, perché cresce maggiormente prima di esso, ad indicare un maggiore apporto della selezione.

In generale usare questi grandi dataset online permette di studiare questi fenomeni, ma i dati ottenuti sono in realtà la media del grande numero di valori. Ciò permette di capire che la differenza tra selezione ed influenza sociale esiste nel formare l'homophily, ma che essa è troppo sottile per essere notata in realtà minori. Inoltre bisognerebbe chiedersi se i risultati ottenuti siano simili o meno a seconda del dominio studiato.

## 4.5 A Spatial Model of Segregation

Uno degli effetti dell'homophily più facili da percepire e notare è la creazione di vicinati che raggruppano persone della stessa razza ed etnia nelle città. Le persone tendono a vivere vicino a persone a loro simili, e di conseguenza aprono negozi, ristoranti e servizi orientati a quelli come loro. Un esempio è quello illustrato da Möbius e Rosenblat nella figura 36, che rappresenta la percentuale di Americani Africani in ogni blocco di Chicago negli gli anni '40 e '60.



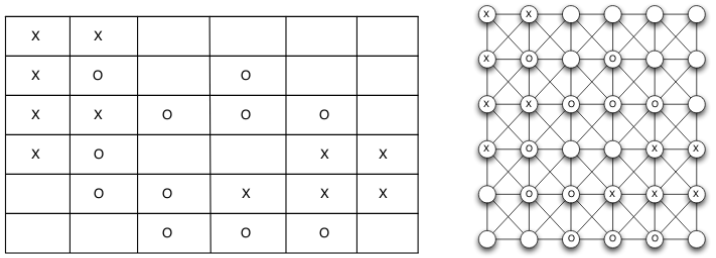
(a) *Chicago, 1940* (b) *Chicago, 1960*

Figura 36: La rappresentazione della segregazione degli Americani Africani negli anni (a) 1940 e (b) 1960 a Chicago.

### 4.5.1 The Schelling Model

Un famoso modello introdotto da Thomas Schelling mostra come dei pattern globali di segregazione spaziale possano nascere dagli effetti dell'homophily che operano a livello locale, concentrandosi in particolare sul fatto che questi effetti possano operare anche quando nessun individuo cerchi direttamente la segregazione.

Il modello è formulato del seguente modo: pensiamo ad un popolazione di individui, che chiamiamo **agents**, in cui ogni agente è di tipo $X$ oppure $O$. Il tipo rappresenta una qualsiasi caratteristica che è alla base dell'homophily. Gli agenti sono situati all'interno di una griglia, come una rappresentazione geografica in due dimensioni di una città. I vicini di una cella sono le celle a contatto con essa, incluse quelle diagonali. Possiamo pensare a questa rappresentazione anche sotto forma di grafo, in cui ogni cella è un nodo e vi è un arco tra due nodi se le corrispettive celle sono vicine.




(a) *Agents occupying cells on a grid.* (b) *Neighbor relations as a graph.*

Figura 37: La rappresentazione del modello di Schelling sotto forma di (a) griglia e di (b) grafo.

Il vincolo fondamentale del modello è che ogni agente vuole avere almeno una parte dei vicini che siano simili a lui. Rappresentiamo questo valore come la soglia $t$, comune a tutti gli agenti. Se un agente nota che ci sono meno vicini simili a lui rispetto alla soglia $t$, egli sarà **unsatisfied** ed avrà il desiderio di spostarsi in una cella in cui questa soglia sia rispettata, come nel caso della figura 38(a).

### 4.5.2 The Dynamics of Movement

Discutiamo ora negli aspetti dinamici del modello, ovvero come gli agenti si possano muovere per essere soddisfatti. Inizialmente si cercano tutti gli agenti non soddisfatti, poi in una sequenza di **rounds** essi vengono spostati uno per volta in una cella in cui essi sono soddisfatti. Dopo aver completato tutte le mosse, troviamo una nuova situazione in cui altri agenti possono essere insoddisfatti; a questo punto comincia un altro round.

Non ci sono regole precise per il movimento: gli agenti possono essere spostati in ordine casuale, oppure seguendo la riga della griglia; possono essere spostati nella cella più vicina, oppure in una casuale tra quelle che lo soddisfano. Nel caso in cui non ci sia una cella che soddisfa l'agente, questo può essere lasciato dove si trova oppure spostato in una cella casuale.

La figura 38(b) mostra il risultato di un round di movimenti, usando come soglia $t$ uguale a 3. Si noti che il livello di segregazione dopo il movimento aumenta.

| | | | | | |
|---|---|---|---|---|---|
| X1* | X2* | | | | |
| X3 | O1* | | O2 | | |
| X4 | X5 | O3 | O4 | O5* | |
| X6* | O6 | | | X7 | X8 |
| | O7 | O8 | X9* | X10 | X11 |
| | | O9 | O10 | O11* | |

(a) *An initial configuration.*

| | | | | | |
|---|---|---|---|---|---|
| | | | | | |
| X3 | X6 | O1 | O2 | | |
| X4 | X5 | O3 | O4 | | |
| | O6 | X2 | X1 | X7 | X8 |
| O11 | O7 | O8 | X9 | X10 | X11 |
| | O5 | O9 | O10* | | |

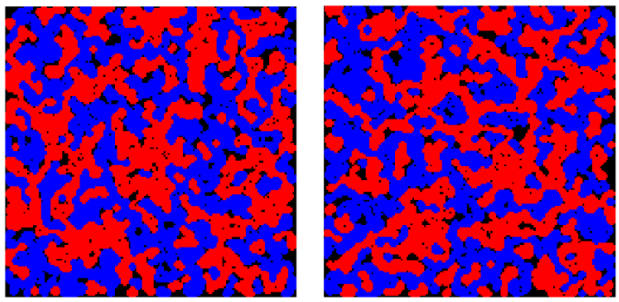
(b) *After one round of movement.*

Figura 38: Dopo aver sistemato gli agenti nelle celle, (a) si trovano gli agenti insoddisfatti e (b) si spostano uno per volta in un luogo in cui sono soddisfatti.

### 4.5.3 Larger Examples

Gli esempi minori come quello visto in precedenza sono utili per capire le basi del modello, ma per ottenere dati significativi ottiene usare moli maggiori di dati. Per questo sono state create delle simulazioni computerizzate per poter identificare i pattern presenti.

La figura 39 mostra i risultati della simulazione scritta da Luke, utilizzando una griglia 150x150, con 10000 agenti, 2500 celle vuote e $t$ pari a 3. Notiamo che in entrambe le simulazioni, in cui cambia la configurazione iniziale, si raggiunge un equilibrio dopo circa 50 rounds, e il posizionamento degli agenti nel cercare la soddisfazione crea delle larghe regioni omogenee, congiunte tra loro.



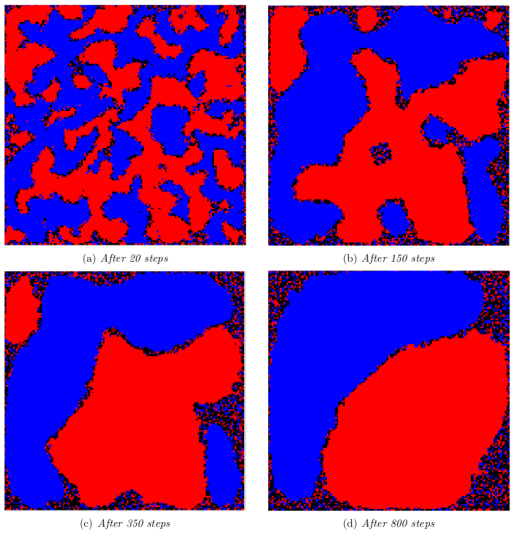
(a) *A simulation with threshold 3.* (b) *Another simulation with threshold 3.*

Figura 39: Due simulazioni del modello di Schelling con $t = 3$.

### 4.5.4 Interpretations of the Model

Avendo visto i risultati delle simulazioni, possiamo dedurre che la segregazione viene attuata anche quando nessun agente cerca attivamente di segregarsi. Tipicamente un agente si attacca ad un cluster con altri simili, e così fanno gli altri: a lungo termine questo porta alla segregazione.

Ciò avviene ancora più radicalmente se si alza la soglia $t$ a 4. Tenendo gli altri parametri come la simulazione precedente, si nota nella figura 40 che dopo molti rounds si arriva ad una situazione in cui è presente solamente una grande regione per tipo.



(a) *After 20 steps* (b) *After 150 steps*

(c) *After 350 steps* (d) *After 800 steps*

Figura 40: La simulazione del modello di Schelling dopo (a) 20, (b) 150, (c) 350 e (d) 800 rounds, con $t = 4$.

Possiamo allora affermare che il modello di Schelling sia un valido esempio per mostrare come componenti immutabili quali razza ed etnia, agendo a livello locale e legandosi a caratteristiche mutabili quali la scelta di dove vivere, possano avere effetti globali.

# 13 The Structure of the Web

Passiamo ora a considerare delle reti dove i nodi non sono rappresentati da individui, bensì da informazioni, legate tra loro in diversi modi. Chiamiamo questo tipo di rete **information network**. L'esempio più grande e famoso di information network è sicuramente il World Wide Web.

Sebbene ci siano molte differenze con le social networks studiate in precedenza, vedremo che sono presenti anche diversi aspetti in comune con esse: useremo gli stessi concetti derivanti dalla teoria dei grafi, come distanze e giant components.

## 13.1 The World Wide Web

A livello base il Web è un'applicazione sviluppata per permettere di condividere informazioni su Internet; è stato creato da Tim Berners-Lee tra l'89 ed il '91. Possiamo evidenziarne due aspetti principali: il primo è che esso offre la possibilità di condividere e rendere disponibile facilmente a tutti dei documenti, chiamati **web pages**, che è possibile creare e rendere disponibili tramite il proprio computer; il secondo è che offre un meccanismo per accedere a questi documenti, chiamato **browser**, che permette di connettersi ai computer in Internet e ottenere questi documenti.

### 13.1.1 Hypertext

Oltre a queste caratteristiche di base, c'è la possibilità di aggiungere in una parte del documento un collegamento virtuale ad un'altra pagina, permettendo al lettore di raggiungerla direttamente. L'insieme delle pagine diventa così un grafo, in cui i nodi sono le pagine stesse e gli archi sono i collegamenti che portano da una pagina all'altra.

La decisione di usare questo tipo di riferimento per costruire il Web è un'applicazione di una creazione di documenti assistita dal computer conosciuta come **hypertext**. Essa permette di sostituire la tradizionale struttura lineare di un testo con una struttura a rete, in cui ogni parte di testo può riferirsi ad ogni altra parte.

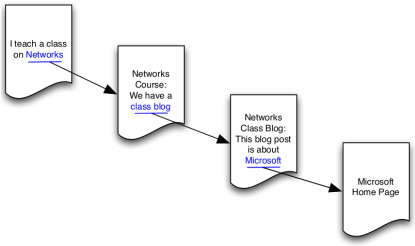



Figura 41: Un esempio di pagine Web e hypertext rappresentati come un grafo.

## 13.2 Information Networks, Hypertext, and Associative Memory

La struttura ipertestuale del Web ci offre un importante e familiare esempio di information network: i nodi contengono informazioni, con i collegamenti espliciti che esprimono la relazione tra i nodi.

### 13.2.1 Intellectual Precursors of Hypertext

Il primo grande precursore dell'ipertesto è il concetto di **citation** tra articoli e libri scolastici. Quando gli autori di un lavoro vogliono creare un contesto per approfondire un'idea, spesso fanno riferimento. Ad esempio la figura 42 mostra la rete di citazioni tra papers di sociologia alla base di alcuni concetti mostrati in precedenza.

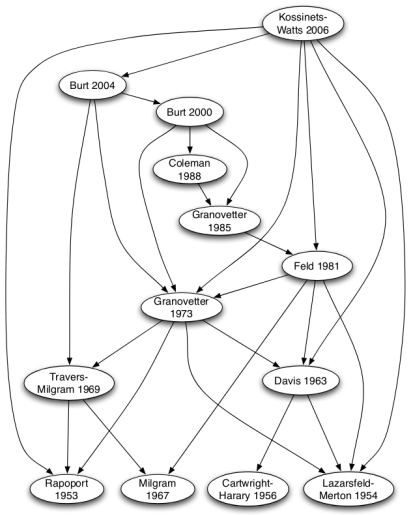



Figura 42: La rete di citazioni tra alcuni papers di ricerca.

Una delle differenze principali tra la rete di citazione ed il Web è l'importanza dell'ambito temporale: è infatti impossibile che se un nodo $X$ fa riferimento ad un altro nodo $Y$, esista anche un arco in senso inverso, in quanto al tempo in cui il nodo $Y$ viene formulato, $X$ non può esistere. Nel Web invece le pagine sono in continua evoluzione e possono modificarsi aggiungendo collegamenti verso pagine più recenti. Sebbene i collegamenti nel Web siano orientati, non c'è un senso di flusso come quello dato dalla linea temporale nelle citation networks.

Un altro tipo di information network è quella che si ottiene dall'associazione di idee a diversi concetti, come il vagare della mente in uno *stream-of-consciousness*. Questa rete è chiamata **semantic network** ed i suoi nodi rappresentano i concetti mentre i collegamenti rappresentano una relazione logica o percepita tra essi.

### 13.2.2 Vannevar Bush and the Memex

Le information networks risalgono a periodi molto precedenti della storia; per secoli esse erano associate alle biblioteche ed alla letteratura scolastica, invece che alla tecnologia e ad Internet come oggi. L'idea di un loro lato tecnologico è accreditata a Vannevar Bush ed ad un suo articolo del 1945, in cui l'autore immaginava in cui le nuove e nascenti tecnologie potessero rivoluzionare il modo in cui si memorizza, scambia ed accede all'informazione.

In particolare Bush ha osservato che tradizionalmente il modo in cui si memorizzano le informazioni è **lineare**, come una collezione di oggetti in un certo ordine. D'altra parte invece il nostro modo di ragionare è legato alla **memoria associativa**, come quella rappresentata dalle semantic networks. Bush ha quindi immaginato un prototipo chiamato **Memex** che funzionasse in maniera simile al Web, contenendo la conoscenza umana in maniera digitale, connessa con link associativi.

L'articolo di Bush prevedeva non solo il Web, ma anche il modo in cui oggi ci riferiamo ad esso, come una enciclopedia universale, un sistema socioeconomico, un cervello globale. Non a caso i primi sistemi di hypertext facevano esplicito riferimento alle sue idee.

### 13.2.3 The Web and Its Evolution

Questi sistemi ci riportano ai primi anni '90, quando le pagine erano per lo più statiche e la maggior parte dei collegamenti avevano la funzione principale di **navigazione**, per portare da una pagina all'altra.

Questa è ancora oggi una approssimazione di buona parte del Web, ma la restante porzione si è evoluta superando questo modello. Questo perché

la capacità computazione è aumentata esponenzialmente negli ultimi anni, permettendo ai server di ospitare ed offrire anche programmi complessi, come ad esempio dei collegamenti del tipo "Aggiungi al carrello" o "Carica l'immagine". Questi collegamenti non hanno lo scopo principale di navigare, ma quello di attivare delle transazioni sulla macchina che ospita il sito.

Possiamo quindi oggi distinguere due tipi di collegamenti, quelli **navigational** e quelli **transactional**. Questa non è una divisione perfetta perché molti link offrono entrambe le funzioni, ma è una dicotomia da tener presente nell'affrontare l'argomento.

Ci concentreremo sulla parte di navigazione del Web, che ne rappresenta la spina dorsale, perciò è importante distinguere i collegamenti che ci interessano, ma per fortuna questo ambito è molto sviluppato, soprattutto nei motori di ricerca, che filtrano i collegamenti transazionali per migliorare la navigazione dell'utente.

## 13.3 The Web as a Directed Graph

Quando guardiamo il Web come un grafo possiamo capire meglio le relazioni logiche espresse dai suoi collegamenti, dividere la sua struttura in componenti minori e trovare le pagine più importanti.

Usando solo la parte di navigazione del Web, è importante sottolineare che i collegamenti sottostanti sono tutti orientati, dal momento che partono da un nodo e arrivano ad un altro. Questa è una differenza importante con le social e le information networks.

### 13.3.1 Paths and Strong Connectivity

Possiamo trovare le componenti connesse anche in un grafo orientato, ma prima dobbiamo ridefinire il concetto di cammino: il cammino da un nodo A ad un nodo B di un grafo orientato è una sequenza di nodi che comincia con A e termina con B, in cui ogni coppia consecutiva è connessa da un arco che punta in avanti.

La figura 43 mostra un esempio di grafo orientato formato da un insieme di pagine Web. Notiamo come il collegamento tra due pagine non sia reciproco, ma orientato.

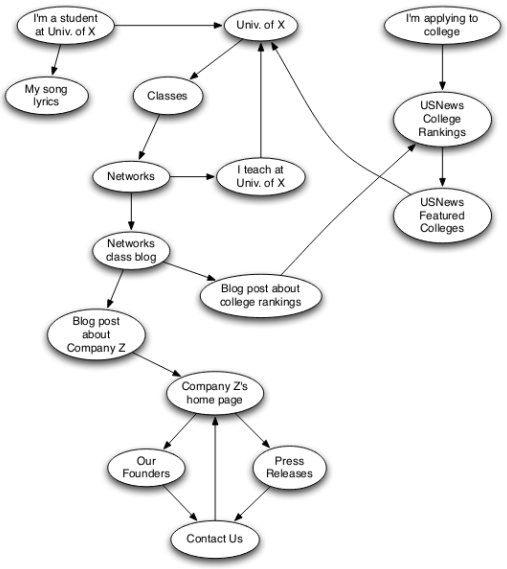



Figura 43: Un esempio di grafo orientato formato da un insieme di pagine Web.

Possiamo ora definire la connessione nei grafi orientati. Diciamo che un grafo orientato è **strongly connected** se c'è un cammino da ogni nodo verso ogni nodo. Il grafo della figura 43 non è fortemente connesso.

### 13.3.2 Strongly Connected Components

Quando un grafo orientato non è fortemente connesso, è importante definire la sua proprietà di **reachability**, che indica quali nodi sono raggiungibili da quali altri. La chiave per descriverla è trovare la nozione di **component**, simile a quella formulata per i grafi non orientati.

> Diciamo che una **strongly connected component** (SCC) in un grafo orientato è il sottoinsieme dei nodi in modo che (i) ogni nodo nel sottoinsieme ha un cammino verso ogni altro nodo e (ii) il sottoinsieme non è parte di un insieme più grande con la proprietà che ogni nodo possa raggiungere ogni un altro.

La figura 44 mostra le componenti fortemente connesse del grafo della figura 43. Guardando l'immagine, possiamo trovare la reachability tra due nodi A e B controllando la SCC in cui si trovano. Se sono entrambi nella stessa, possono raggiungersi. Altrimenti bisogna guardare le SCCs come "supernodi" e vedere se c'è un cammino che collega la componente di A con quella di B nella dritta direzione. Se questi collegamenti non esistono, non c'è un cammino tra A e B.

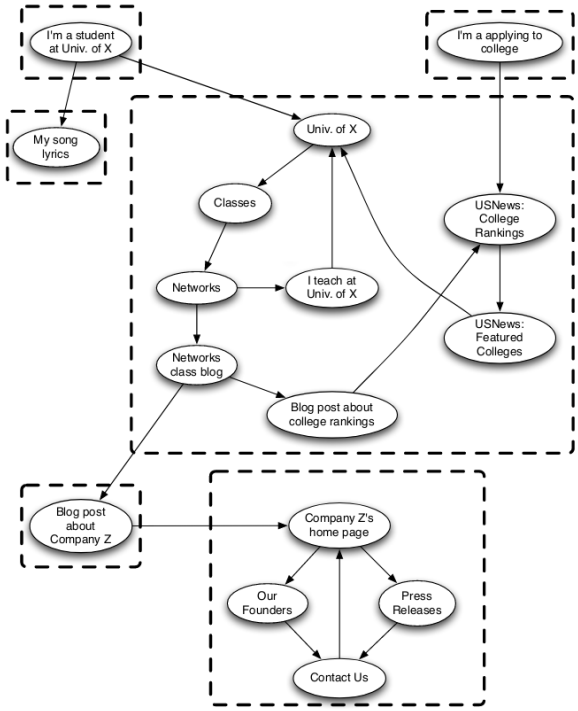



Figura 44: Le componenti fortemente connesse di grafo orientato formato da un insieme di pagine Web.

## 13.4 The Bow-Tie Structure of the Web

Nel 1999, dopo una grande crescita del Web, Andre Broder ed i suoi colleghi decisero di creare una mappa mondiale del Web, usando le componenti fortemente connesse come dei blocchi per costruirla. Per fare ciò usarono di dati di AltaVista, il più diffuso motore di ricerca del tempo.

### 13.4.1 A Giant Strongly Connected Component

Ciò che Broder et al. volevano trovare era qualcosa di più concettuale, dividendo il Web in diversi pezzi e mostrando come questi siano legati tra loro.

La loro prima scoperta fu che il Web contenesse una componente fortemente connessa gigante, contenente la maggior parte dei nodi. Questo perché molte pagine sono legate tra loro grazie ai motori di ricerca e perché molte delle pagine principali contengono dei collegamenti con le altre pagine principali, che siano governative, commerciali o di organizzazioni non-profit.

### 13.4.2 The Bow-Tie Structure

La parte rimanente dello studio era collegare le SCCs rimanente con quella gigante, classificando i nodi ed il loro collegamenti con quelli contenuti nella SSC gigante. Gli insiemi trovati sono:

1. la componente *IN*, che contiene i nodi che possono raggiungere la SCC gigante ma che non possono essere raggiunti da essa;

2. la componente *OUT*, che contiene i nodi che possono essere raggiunti dalla SCC gigante ma che non possono raggiungerla;

3. la componente *tendrils*, che contiene i nodi che possono essere raggiunti da IN ma non possono raggiungere la SCC gigante, ed i nodi che raggiungono OUT ma non possono essere raggiunti dalla SCC gigante;

4. l'insieme *disconnected*, che comprende tutti i nodi che non avrebbero un cammino verso la SCC gigante anche trasformando il grafo in non orientato; essi non fanno parte di nessuna delle precedenti categorie.

Come vediamo nella figura 45 il Web prende la forma di un fiocco; grazie a questa visione abbiamo una vista ad alto livello della struttura del Web, basata sulla proprietà di reachability e su come le componenti si leghino tra loro. Questa visione però non offre informazioni sulle connessioni a livello più basso, ma di questo ci occuperemo più avanti.

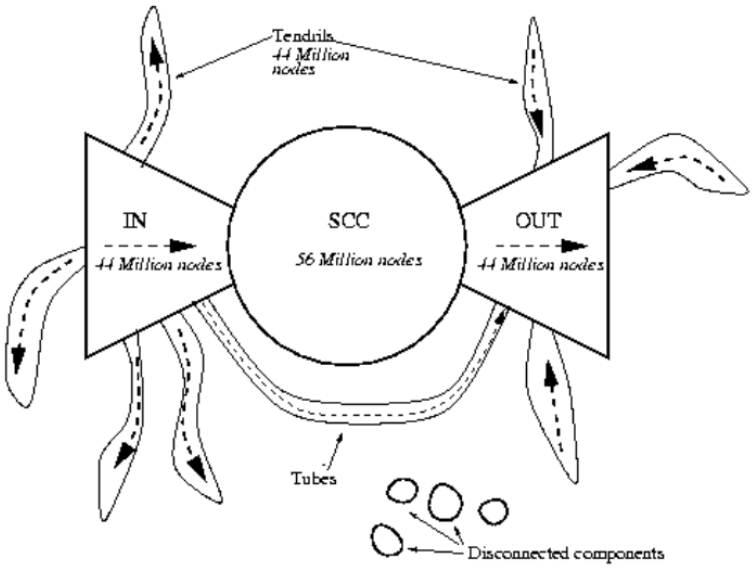



Figura 45: Una visione schematica della struttura a fiocco del Web.

## 13.5 The Emergence of Web 2.0

La crescente ricchezza dei contenuti del Web ha portato a significativi cambiamenti nella sua struttura nel primo decennio degli anni 2000. Le tre principali forze dietro questi cambiamenti furono:

1. la crescita degli authoring styles che ha aumentato i contenuti prodotti dagli utenti e permesso di creare a mantenere diverso contenuto condiviso;

2. lo spostamento dei dati privati dai computer personali ai servizi offerti online ospitati dalle grandi compagnie;

3. la crescita degli stili dei link, ad enfatizzare non solo i collegamenti tra documenti ma anche tra persone.

Messi insieme, questi ambiti hanno portato Tim O'Really and altri esperti a parlare della "emergenza del Web 2.0", affermando che esso sia più un'attitudine e non una tecnologia. Nonostante ciò, proprio in quel periodo di vede la nascita di Wikipedia che incentiva gli utenti a collaborare tra loro, Gmail che porta l'utente ad affidare loro i propri dati, MySpace e Facebook come idea di rete sociale online. Possiamo anche citare i servizi di Flickr, YouTube e Twitter.

I curatori dei siti si sono inoltre sempre più interessati al feedback degli utenti, e ciò spiega come alcuni principi che hanno portato all'emergenza del web 2.0 siano in realtà il risultato di fenomeni sociali, come:

- "il software migliora se usato da più persone", molti siti infatti migliorano se le persone si interessano ad esso;

- "la saggezza della massa", come ad esempio il caso di Wikipedia;

- "la lunga coda", come spiega come in realtà ci sia un equilibrio tra un numero ridotto di siti popolari e una "long tail" di contenuti di nicchia specifici.

Inoltre possiamo citare i sistemi di "trust" o "reputation" che permettono di segnalare il comportamento degli altri utenti, ed i sistemi di "recommendation" che guidano gli utenti verso ciò che conoscono poco o nulla, distribuendo la popolarità e la long tail.

# 5 Positive and Negative Relationships

Fino ad ora abbiamo trattato le relazioni come positive: relazioni di amicizia, collaborazione, condivisione di informazioni o appartenenza ad un gruppo. Ma in molte reti ci sono anche effetti negativi in gioco, come inimicizie, disaccordi e conflitti.

La nozione di **structural balance** è ciò permette di analizzare e studiare la tensione tra i collegamenti positivi e quelli negativi. Inoltre possiamo vedere come queste tensioni a livello locale possano portare a degli effetti a livello globale nella rete.

## 5.1 Structural Balance

Supponiamo di avere una rete sociale in cui tutti gli individui conoscono tutti gli altri, ossia vi è un collegamento tra tutti i nodi. Questo tipo di rete è chiamato **clique** o **complete graph**. Possiamo **etichettare** ogni arco con un + o un - per indicare rispettivamente una relazione positiva o negativa. Dal momento che il grafo è completo sappiamo che c'è una relazione tra ogni coppia di nodi, perciò questo modello rispecchia quella che può essere la situazione in un piccolo gruppo di persone, come ad esempio in una classe scolastica, in cui tutti bene o male si conoscono tra loro.

L'idea principale alla base dello structural balance risale agli studi sulla psicologia di Heider negli anni '40, ampliati poi nell'ambito dei grafi col lavoro di Cartwright e Harary negli anni '50. Essa può essere formulata nel seguente modo: se prendiamo due individui qualsiasi, l'arco tra loro può essere positivo o negativo, ma se prendiamo un gruppo di tre persone, certe configurazioni di + e - sono psicologicamente e socialmente più plausibili di altre.

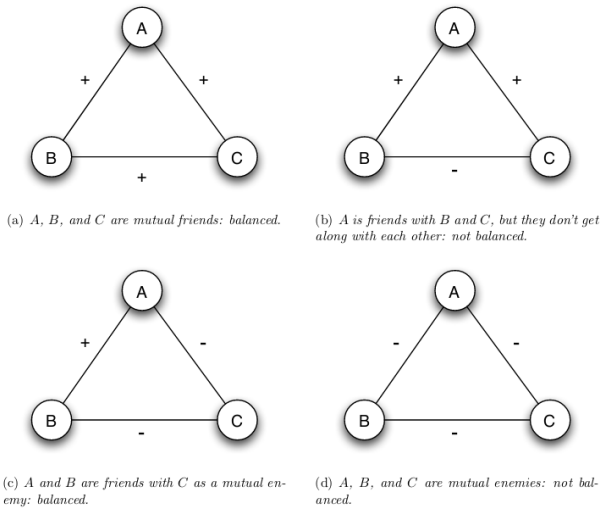
La figura 46 mostra le quattro configurazioni possibili tra tre nodi A,B e C:

1. se tutti gli archi sono positivi (a), è una situazione naturale in cui tutti sono amici ed è bilanciata;

2. se vi è una sola relazione positiva e due negative (c), è una situazione naturale in cui due sono amici ed hanno un nemico in comune;

3. se vi sono due relazioni positive ed una sola negativa (b), si tratta di una relazione instabile, poiché un individuo è amico di altri due, che però non vanno d'accordo tra loro

4. se tutti gli archi sono negativi (d), vi è instabilità perché potrebbe esserci la tendenza da parte di due individui di allearsi contro l'altro.




(a) *A, B, and C are mutual friends: balanced.*

(b) *A is friends with B and C, but they don't get along with each other: not balanced.*

(c) *A and B are friends with C as a mutual enemy: balanced.*

(d) *A, B, and C are mutual enemies: not balanced.*

Figura 46: Le quattro possibili configurazioni di balance tra tre nodi.

In base a queste osservazioni, un triangolo è bilanciato se formato da un numero dispari di relazioni positive.

### 5.1.1 Defining Structural Balance for Networks

Finora abbiamo parlato di structural balance per gruppi di tre nodi, ora cerchiamo di generalizzare questa nozione in modo che si possa applicare ad un insieme più vasto, come l'intera rete.

In particolare, diciamo che un grafo completo etichettato è bilanciato se ogni triangolo in esso è bilanciato, ovvero se:

> Structural Balance Property: per *ogni* insieme di tre nodi, tutti e tre gli archi che li uniscono sono positivi, oppure solamente uno è positivo.

Per esempio nella seguente figura 47 il primo grafo è bilanciato, mentre il secondo no perché nel triangolo A,B,C vi sono due archi positivi.

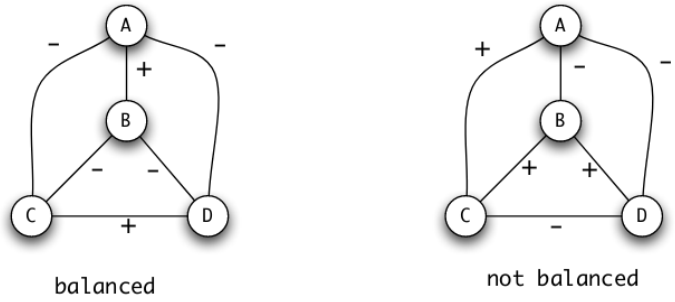



Figura 47: Un esempio di grafo bilanciato ed uno non bilanciato.

La definizione di grafo bilanciato è però un po' estrema, perché difficilmente in situazioni reali si riesce ad eliminare ogni triangolo non bilanciato, perciò si potrebbe semplificare la definizione per considerare solo una determinata percentuale di triangoli bilanciati.

## 5.2 Characterizing the Structure of Balanced Networks

Come è fatta a livello generale una rete bilanciata? Supponiamo di avere un grafo completo bilanciato in cui i nodi possono essere divisi in due gruppi, $X$ e $Y$, tali che ogni nodo in $X$ ed $Y$ ha una relazione positiva con gli altri membri del gruppo, ma ogni elemento in $X$ è nemico di ogni elemento in $Y$ come nella figura 48. Possiamo verificare che questa rete è bilanciata: un triangolo contenuto interamente in un gruppo ha tre relazioni positive, un triangolo con due persone in un gruppo ed una nell'altro ha una sola relazione positiva.

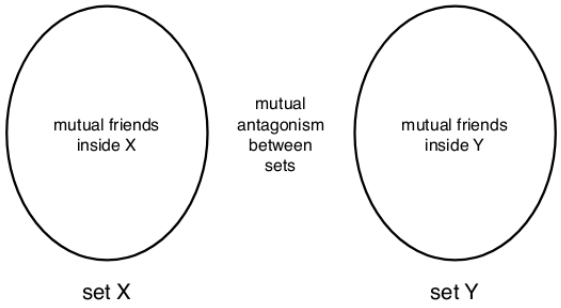



Figura 48: La struttura generale di un grafo bilanciato.

Il **balance theorem**, dimostrato da Frank Harary, afferma che l'unico modo di ottenere un grafo bilanciato è che ci siano solo relazioni positive,

oppure che la struttura sia riconducibile a quella dei due gruppi formulata in precedenza, ed è definito in questo modo:

> The Balance Theorem: se un grafo completo etichettato è bilanciato, allora o tutti i nodi sono amici, oppure i nodi possono essere divisi in due gruppi, $X$ e $Y$, tali che ogni coppia di nodi in $X$ è amica, ogni coppia di nodi in $Y$ è amica, ed ogni nodo in $X$ è nemico di ogni nodo in $Y$.

Stiamo quindi prendendo la proprietà locale di structural balance e mostrando come questa implichi una proprietà globale: o tutti sono amici, o vi sono due grandi gruppi.

### 5.2.1 Proving the Balance Theorem

Per dimostrare il teorema, supponiamo di avere un grafo completo etichettato ed assumiamo che sia bilanciato. Se non ci sono archi negativi, il teorema è dimostrato. In caso contrario dobbiamo dimostrare l'esistenza dei due gruppi citati nel teorema. Prendiamo quindi un nodo A e definiamo l'insieme $X$ come quello composto da A ed i suoi amici e l'insieme $Y$ come quello composto dai nemici di A.

Il teorema deve quindi soddisfare tre condizioni:

1. ogni coppia di nodi in $X$ è amica;
2. ogni coppia di nodi in $Y$ è amica;
3. ogni nodo in $X$ è nemico di ogni nodo in $Y$.

Per la condizione (1), sappiamo che A è amico di ogni nodo in X, perciò se supponiamo che sia amico di B e C, anche B e C devono essere amici tra loro, altrimenti avremmo un triangolo non bilanciato; perciò i nodi in $X$ sono tutti amici. Per la condizione (2), prendiamo due nodi D ed E in $Y$; sappiamo che entrambi sono nemici di A, perciò devono essere amici tra loro, altrimenti avremmo nuovamente un triangolo non bilanciato: tutti i nodi in $Y$ sono amici tra loro. Infine per la condizione (3) prendiamo i nodi B e D, sappiamo che A è amico di B e nemico di D, quindi B e D devono essere nemici per ottenere un triangolo bilanciato; otteniamo quindi che i nodi in $X$ sono nemici dei nodi in $Y$. Questo completa la dimostrazione del teorema; la figura 49 mostra lo schema su cui essa è basata.

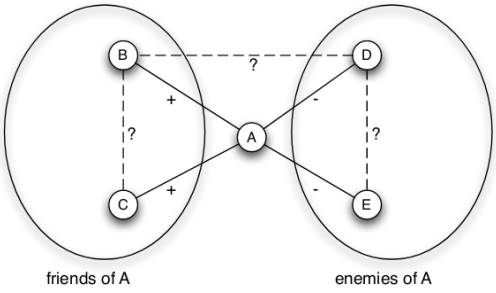


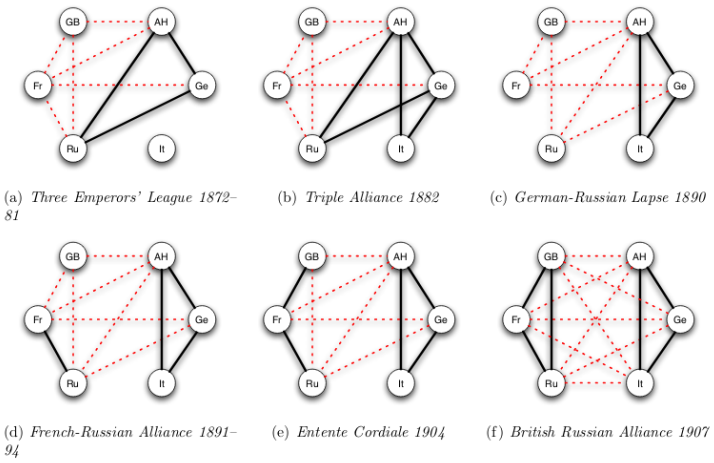

Figura 49: Lo schema dell'analisi di una rete bilanciata.

## 5.3 Applications of Structural Balance

Alcune ricerche recenti si sono interessate degli aspetti dinamici dello structural balance, modellando come amicizie e antagonismi possono evolversi nel tempo all'interno di una rete, alla ricerca di un equilibrio. Antal, Krapivsky e Redner studiarono un modello in cui si comincia con un assegnamento casuale delle etichette, poi si cercano i triangoli non bilanciati e si cambia una etichetta per bilanciarli. Questo procedimento simula una situazione in cui le persone cambiano continuamente amici e nemici alla ricerca di un equilibrio, ma le nozioni matematiche alla base del processo sono molto complesse.

### 5.3.1 International Relations

La politica internazionale è un contesto in cui è naturale ritrovare una situazione simile a quella descritta in precedenza, descrivendo un grafo in cui i nodi sono le nazioni e gli archi indicano alleanze o animosità. Uno degli esempi usati da Antal, Krapivsky e Redner è il cambiamento delle alleanze nel periodo precedente la prima guerra mondiale, mostrato nella figura 50. Questo mostra che l'equilibrio non è per forza qualcosa di positivo, in quanto può portare ad una opposizione difficile da risolvere tra due parti.




(a) *Three Emperors' League 1872–81* (b) *Triple Alliance 1882* (c) *German-Russian Lapse 1890*

(d) *French-Russian Alliance 1891–94* (e) *Entente Cordiale 1904* (f) *British Russian Alliance 1907*

Figura 50: L'evoluzione delle alleanze in Europa nel periodo 1872-1907.

### 5.3.2 Trust, Distrust, and Online Ratings

Una grande fonte di dati di rete con archi positivi e negativi proviene direttamente dalle comunità di utenti sul Web. Ad esempio Epinions è un sito che permette di valutare diversi prodotti ed esprimere fiducia o diffidenza verso altri utenti.

Guha et al. studiarono proprio i dati raccolti da questo sito per mostrare similitudini e differenze con la dicotomia amico - nemico nella teoria di structural balance. Una delle differenze principali è il fatto che nello studio dello structural balance abbiamo considerato il grafo come non orientato, mentre qui esso è orientato, perché non è detto che un utente venga corrisposto nella valutazione di un altro.

Inoltre nel caso in cui A diffidi di B, e B diffidi di C, che legame dovremmo aspettarci da parte di A verso C? Dipende essenzialmente dal contesto, possiamo vedere la situazione come structural balance e quindi che A si fidi di C; oppure se pensiamo che A diffidi di B perché si sente più competente, e B a sua volta faccia lo stesso con C, allora anche A dovrebbe diffidare di C anche più di quanto espresso verso B.

Capire come funzionino queste relazioni positive e negative è importante per capire il loro ruolo nel Web sociale dove gli utenti si valutano tra loro. La ricerca in questo campo può portare a rispondere ai dubbi prima citati e capire come le teorie di equilibrio siano applicate nei grandi datasets.

## 5.4 A Weaker Form of Structural Balance

Studiando le relazioni positive e negative sono state formulate delle definizioni alternative di structural balance. Ad esempio James Davis afferma che nel caso di due relazioni positive in una triangolo, vi è la possibilità che avendo un amico in comune i due nemici cerchino di riconciliarsi; oppure egli dice che nel caso di un triangolo con tre relazioni negative, è difficile che vi sia la possibilità che una coppia si riappacifichi. Questo va quindi contro le assunzioni fatte in precedenza per definire lo structural balance.

### 5.4.1 Characterizing Weakly Balanced Networks

Più precisamente diciamo che un grafo completo etichettato è **weakly balanced** se vale la seguente proprietà:

> Weak Structural Balance Property: non c'è un insieme di tre nodi formato esattamente da esattamente due archi positivi ed uno negativo.

Dato che questa proprietà è meno restrittiva della definizione precedente, ci possiamo aspettare un insieme più ampio di strutture possibili in una rete debolmente bilanciata. Come mostra la figura 51 possono nascere più insiemi che raggruppano tutti gli elementi amici tra loro, ma nemici di quelli in altri gruppi.

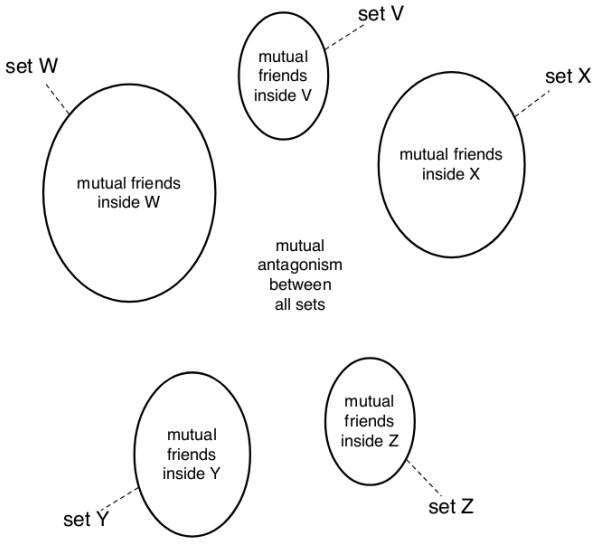



Figura 51: Una possibile struttura di un grafo completo debolmente bilanciato.

Possiamo facilmente dimostrare che una struttura del genere è debolmente bilanciata: in ogni triangolo con due archi positivi, il terzo nodo deve per forza essere compreso nello stesso gruppo e l'arco relativo è quindi positivo anch'esso.

Come per il teorema di structual balance, possiamo formulare la proprietà per un grafo debolmente bilanciato che abbia qualsiasi numero di insiemi:

> Characterization of Weakly Balanced Networks: se un grafo completo etichettato è debolmente bilanciato, allora i suoi nodi possono essere divisi in gruppi in modo che ogni coppia di nodi appartenenti allo stesso gruppo sia amica ed ogni coppia di nodi appartenenti a gruppi diversi sia nemica.

La nozione di weak structural balance permette quindi di andare oltre la definizione di Cartwright-Harary e di ragionare su reti divise in più di due gruppi.

### 5.4.2 Proving the Characterization

Non è difficile dimostrare questa proprietà, modificando in parte la dimostrazione vista in precedenza per la structural balance. Avendo un grafo completo debolmente bilanciato, prendiamo un nodo A e chiamiamo $X$ l'insieme dei suoi amici, ponendo due condizioni:

1. tutti gli amici di A sono amici tra loro
2. A e tutti i suoi amici sono nemici di ogni altro nodo nel grafo.

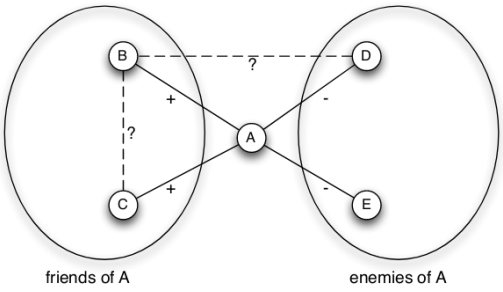



Figura 52: Uno schema dell'analisi della proprietà di bilanciamento debole.

Con l'aiuto della figura 52, per la condizione (1) consideriamo due nodi B e C, entrambi amici di A. Se fossero nemici tra loro, in triangolo ABC avrebbe esattamente due relazioni positive, perciò non è possibile. Per la

condizione (2) consideriamo un nodo D all'infuori di $X$, nemico quindi di A; se B e D fossero amici il triangolo ABD avrebbe nuovamente due archi positivi, cosa non possibile. B e D devono quindi essere nemici.

Dato che valgono entrambe le condizioni, possiamo considerare l'insieme $X$ come il primo gruppo e rimuoverlo dal grafo. Ripetiamo lo stesso procedimento con tutti gli altri gruppi, e dato grazie alle due proprietà dimostriamo la caratterizzazione.

Notiamo che in questa dimostrazione non è servito considerare l'arco tra D ed E, perché il bilanciamento debole non impone condizioni su di esso.

## 5.5 Advanced Material: Generalizing the Definition of Structural Balance

In questa sezione vedremo dei modi più generali per formulare lo structural balance in un rete, in particolare togliendo la condizione di usare un grafo completo e contemplando solamente una percentuale di triangoli bilanciati, invece che tutti.

### 5.5.1 Structural Balance in Arbitrary (Noncomplete) Networks

Consideriamo prima il caso in cui la rete non sia necessariamente completa, ma che ogni arco sia comunque etichettato, come nella figura 53.

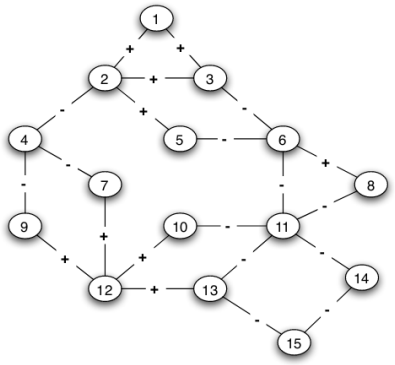



Figura 53: Un grafo non completo ma etichettato.

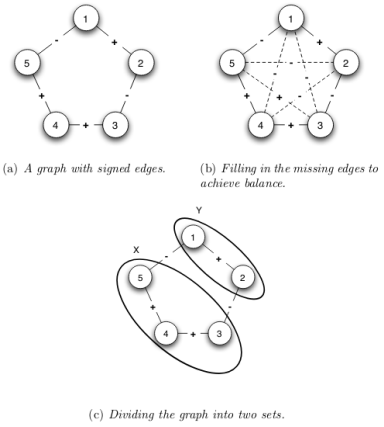
**Defining Balance for General Networks.** Sappiamo che lo structural balance può essere visto a livello locale di ogni triangolo, oppure come proprietà globale:

1. una possibilità sarebbe quella di considerare il problema delle reti non complete come il dover "completare i valori mancanti", ovvero che tutti

abbiano una opinione sugli altri ma non sia stata osservata. Un grafo sarebbe quindi bilanciato nel caso in cui fosse possibile completare la sua struttura inserendo gli archi mancanti. La figura 54 (a) mostra un grafo non completo con gli archi etichettati, mentre (b) mostra come gli archi rimanenti possono essere inseriti per ottenere un grafo bilanciato;

2. l'alternativa sarebbe considerare la visione globale dello structural balance come una divisione della rete in due insiemi mutualmente opposti. Potremmo quindi definire un grafo etichettato come bilanciato se fosse possibile dividere i nodi in due gruppi, formati da relazioni positive tra coppie di nodi all'interno dello stesso gruppo e relazioni negative tra coppie di gruppi diversi, come mostrato nella figura 54 (c).

Questi due modi di considerare il bilanciamento sono equivalenti: un grafo è bilanciato per la prima definizione solo se lo è anche per la seconda, e viceversa. Questo suggerisce una certa naturalezza nella definizione, nonostante i diversi modi per arrivarci.




(a) *A graph with signed edges.* (b) *Filling in the missing edges to achieve balance.*

(c) *Dividing the graph into two sets.*

Figura 54: I due modi usati per definire lo structural balance in un grafo arbitrario.

**Characterizing Balance for General Networks.** Concettualmente in ogni caso qualcosa non è soddisfacente in entrambe le definizioni: entrambe non offrono modo di capire velocemente se un grafo sia bilanciato o meno. Ci possono essere più modi per etichettare gli archi mancanti o scegliere gli insiemi, e cosa ci dice che effettivamente un grafo non sia bilanciabile in alcuni casi?

Consideriamo quindi la seguente domanda: cosa impedisce ad un grafo di essere bilanciato? La figura 55 mostra un esempio di ciò: se cominciamo

dal nodo 1 ad assegnare un gruppo ad ogni nodo in senso orario, ogni scelta è obbligata e non arriviamo mai ad una situazione stabile, continuando in un ciclo. La ragione è in realtà semplice: facendo un passo alla volta e trovando un numero dispari di archi negativi si arriva ad una contraddizione nell'assegnamento del nodo 1 stesso, trovando un rafo non bilanciato.

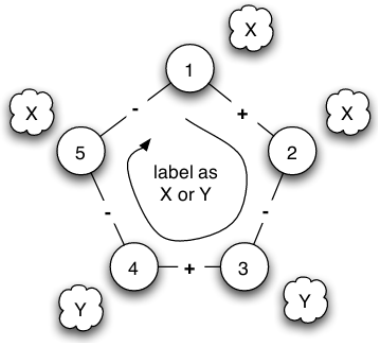



Figura 55: Un grafo non bilanciato, contenente un ciclo con un numero dispari di archi negativi.

Questa regola vale in generale: se il grafo comprende un ciclo con un numero dispari di archi negativi, esso non è bilanciato:

> Claim: un grafo etichettato è bilanciato se e solo se non contiene cicli con un numero dispari di archi negativi.

**Proving the Characterization: Identifying Supernodes.** La procedura di dividere i nodi di un grafo in due insiemi $X$ e $Y$ con le proprietà già più volte citate è chiamata **balanced division**. Vediamo ora una procedura per cercare una balanced division, che o ha successo oppure si ferma trovando un ciclo con un numero di archi negativi dispari.

Il primo step è convertire il grafo in un sottoinsieme contenente solo gli archi negativi, il secondo è di analizzare il problema su questo grafo ridotto. Osservando che i nodi collegati da archi positivi fanno parte dello stesso gruppo, possiamo dividere il grafo in componenti usando questa proprietà. Chiameremo ognuna di queste componenti **supernode**, ed ognuna di esse è connessa da archi positivi al suo interno ed archi negativi verso le altre. Vediamo un esempio nella figura 56.

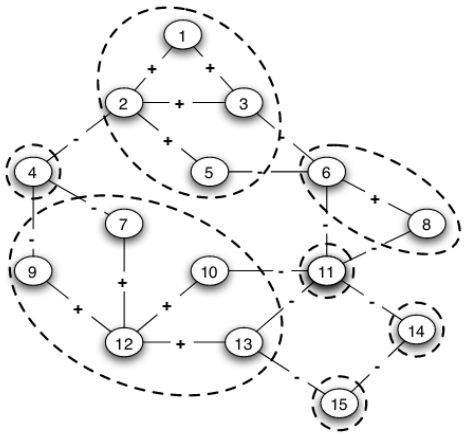



Figura 56: La divisione di un grafo in supernodi.

Ora se ogni supernodo contiene un arco negativo tra due nodi, sappiamo con certezza che quello è un ciclo con numero dispari di archi negativi, perciò il grafo non è bilanciato. Se questo invece non succede, ora possiamo procedere ad etichettare come $X$ o $Y$ ogni supernodo, in modo che ogni scelta sia consistente con le altre. Quindi "collassiamo" ogni supernodo ad un nodo singolo che lo rappresenta, con un arco che lo collega agli altri nodi se i rispettivi supernodi avevano un arco tra loro, ottendo il grafo ridotto della figura 57, in cui è più facile trovare cicli negativi.

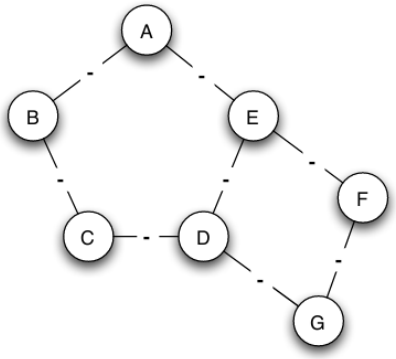



Figura 57: Un grafo ridotto.

**Proving the Characterization: Breadth-First Search of the Reduced Graph.** Questa procedura produce solo due esiti possibili:

1. il primo esito possibile prevede di etichettare ogni nodo nel grafo ridotto come $X$ oppure $Y$ in modo che ogni arco abbia gli endpoints di tipo

diverso. Da questa etichettatura possiamo ricondurci al grafo originale assegnando ad ogni nodo il gruppo del supernodo che lo contiene;

2. la seconda possibilità è quella di trovare un ciclo nel grafo ridotto che abbia un numero dispari di archi. Si può poi convertire questo ciclo in un ciclo del grafo originale contenente un numero dispari di archi negativi; questo connette i supernodi e corrisponde ad un insieme di archi negativi nel grafo originale. Possiamo mettere insieme questi archi negativi usando dei cammini formati da soli archi positivi che passano all'interno di un supernodo, e questo cammino conterrà un numero dispari di archi negativi nel grafo originale. Vediamo un esempio nella figura 58.

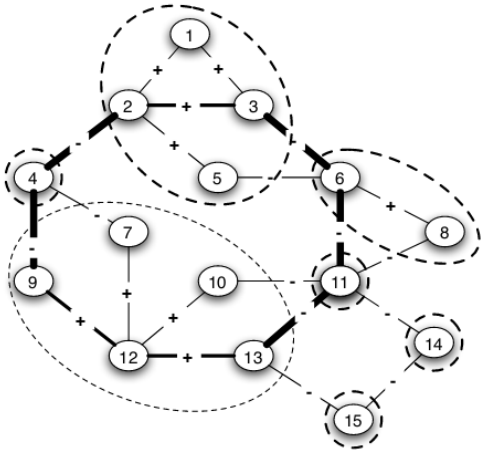



Figura 58: Trovando un ciclo negativo nei supernodi, possiamo estenderlo al grafo originale.

Questa versione del problema in cui il grafo sottostante ha solo archi negativi è noto nella teoria dei grafi come il problema di determinare se il grafo sia **bipartito**, ossia i suoi nodi possono essere divisi in due gruppi (come $X$ e $Y$).

Possiamo quindi attuare una ricerca in ampiezza da ogni nodo "radice" nel grafo, ottenendo dei livelli di nodi ad una certa distanza, come nell'esempio della figura 59, in cui si usa il nodo G come radice e si scende di livello. Ogni arco connette nodi di livelli adiacenti o dello stesso livello; nel primo caso troviamo la divisione dei nodi nei due insiemi assegnando i nodi dei livelli pari ad $X$ e quelli dei livelli dispari ad $Y$. Nel secondo caso, per ogni coppia di nodi A e B dello stesso livello c'è un cammino discendente che porta ad essi; chiamiamo D l'ultimo nodo comune dei due cammini, ed indichiamo con

$k$ la lughezza AD, uguale a BD. Si forma quindi un ciclo formato da questi due cammini più l'arco AB, di lunghezza $2k + 1$, quindi dispari.

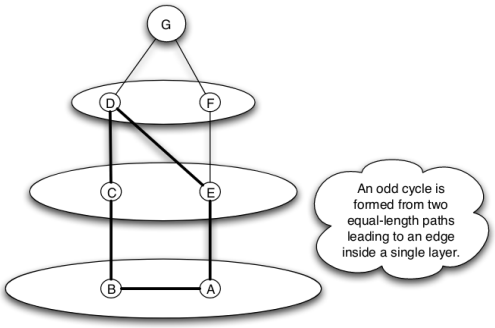



Figura 59: I livelli trovati effettuando la ricerca in ampiezza.

Ricapitolando: se tutti gli archi nel grafo ridotto connetto nodi in livelli adiacenti della ricerca in ampiezza, allora possiamo etichettare i nodi nel grafo ridotto come $X$ o $Y$, avendo anche una divisione bilanciata dei nodi nel grafo originale negli insiemi $X$ e $Y$ ed ottenendo quindi un grafo bilanciato. Altrimenti c'è un arco che connette due nodi nello stesso livello, portando ad un ciclo dispari, convertibile in un ciclo contenente un numero dispari di archi negativi nel grafo originale. Poiché ci sono solo due possibilità, l'affermazione è dimostrata.

### 5.5.2 Approximately Balanced Networks

Torniamo ad occuparci di grafi completi e pensiamo ora a come generalizzare lo structural balance. Le condizioni imposte in precedenza sono abbastanza estreme, perché si richiede che ogni triangolo sia bilanciato. Possiamo quindi rilassare questa condizione e generalizzarla come segue:

> Claim: sia $\varepsilon$ un qualsiasi numero tale che $0 \leq \varepsilon \leq \frac{1}{8}$, e definiamo $\delta = \sqrt[3]{\varepsilon}$. Se almeno 1 - $\varepsilon$ di tutti i triangoli nel grafo completo etichettato sono bilanciati, allora:
>
> 1. c'è un set almeno 1 - $\delta$ nodi in cui almeno 1 - $\delta$ delle coppie sono amiche, oppure
> 2. i nodi possono essere divisi in due gruppi $X$ e $Y$, tali che:
>    (a) almeno 1 - $\delta$ delle coppie in $X$ è amica,
>    (b) almeno 1 - $\delta$ delle coppie in $Y$ è amica, e
>    (c) almeno 1 - $\delta$ delle coppie con un nodo in $X$ ed un nodo in $Y$ sono nemiche.

Il balance theorem formulato in precedenza corrisponde quindi a questa definizione con $\varepsilon = 0$.

Per dimostrare questa formulazione occorre prima contare il numero di archi e triangoli presenti nel grafo; poi si passa alla ricerca di un nodo adatto, ovvero che non sia parte di molti triangoli non bilanciati. Il grafo viene poi diviso secondo questo nodo in due insiemi, uno contenente i suoi amici e l'altro i suoi nemici; data la scelta del nodo, ogni insieme conterrà pochi archi negativi al suo interno. Con queste premesse si può dimostrare la correttezza della formulazione precedente.

# 14 Link Analysis and Web Search

## 14.1 Searching the Web: The Problem of Ranking

Spesso quando si cerca qualcosa online, il primo risultato ottenuto è quello che si cercava. Ma come fanno i motori di ricerca a sapere quale fosse la migliore risposta? "Semplicemente" guardando il Web essi riescono ad assegnare un rank alle pagine, quindi le informazioni necessarie per questo processo sono intrinseche al Web stesso.

Partiamo dal presupposto che la ricerca è un problema difficile nei computer, anche al di fuori dell'ambito del Web, perché le parole chiave spesso forniscono poche informazioni rispetto all'ambito più complesso che si sta cercando, esistono diversi sinonimi per cercare la stessa informazione e la polisemia può rendere difficile capire l'ambito esatto di ricerca.

Con l'arrivo del Web questo problema si ampliò, esplodendo in scala e complessità. Il fatto che si possa creare una pagine su qualsiasi argomento a piacere pone il problema di definire le fonti migliori e più affidabili, mentre una volta chi spendeva risorse e tempo per scrivere un libro su un argomento, generalmente era considerato attendibile. D'altro lato, aumenta anche il numero di richieste per le risorse, ed è difficile interpretare lo scopo della ricerca di un preciso utente, con tutte le accezioni possibili che egli può cercare. Un altro problema importante è il fatto che il contenuto del Web sia dinamico, e che spesso le persone cerchino informazioni in relazione a degli eventi, cambiando per un periodo l'interesse della ricerca. Inoltre si è passati da un problema di mancanza di contenuti, ad un'abbondanza di questi.

## 14.2 Link Analysis using Hubs and Authoritie

Torniamo quindi alla domanda iniziale: cosa permette di capire il risultato migliore per una richiesta?

### 14.2.1 Voting by In-Links

Ad esempio cercando "Cornell", il primo risultato offerto è il sito ufficiale di Cornell, eppure questo al suo interno non cita molte volte la parola chiave cercata. Piuttosto, gli altri siti rilevanti per la ricerca spesso hanno al loro interno un collegamento ad esso.

Questo mostra come i collegamenti siano importanti per classificare le pagine; bisogna però considerare che spesso i collegamenti possono essere fuori argomento, critici o pubblicitari. Per ovviare a questo problema serve prima raccogliere un grande numero di pagine relative alla ricerca; poi si

lascia che le pagine "votino" in base al numero di collegamenti contenuti, ed è per in questo modo che si ottiene il sito più rilevante.

### 14.2.2 A List-Finding Technique

Nel caso invece in cui la parola cercata fosse generica, come "quotidiani"? Tipicamente si otterrà un insieme di risultati contenente i principali quotidiani e alcune pagine che ricevono molti collegamenti in-links come Yahoo!, Facebook o Amazon, assieme ad alcune pagine che contengono una lista di risorse inerenti a quanto richiesto. Vediamo un esempio nella figura 60.

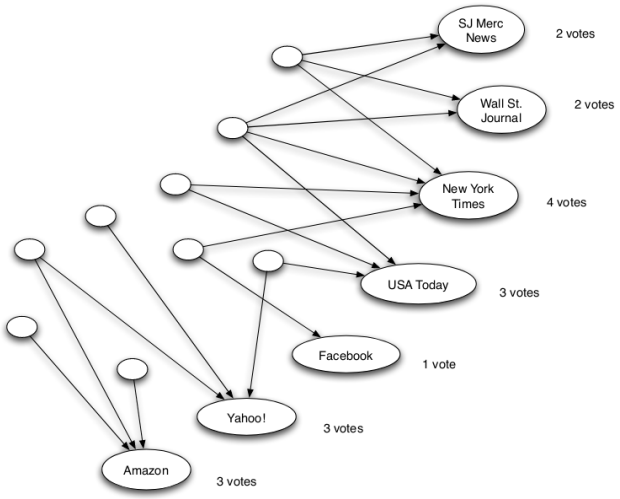



Figura 60: La conta dei collegamenti in-links per la ricerca "quotidiani".

Se osserviamo le pagine votanti, notiamo che alcune hanno votato per molte delle pagine che han ricevuto più voti: questo fa pensare che effettivamente queste pagine abbiano un senso di dove siano le risposte migliori. Possiamo quindi dire che il valore di una pagine come lista è uguale alla somma dei voti ricevuti dalle pagine per cui ha votato: questo permette di trovare le migliori liste, come mostrato nella figura 61.

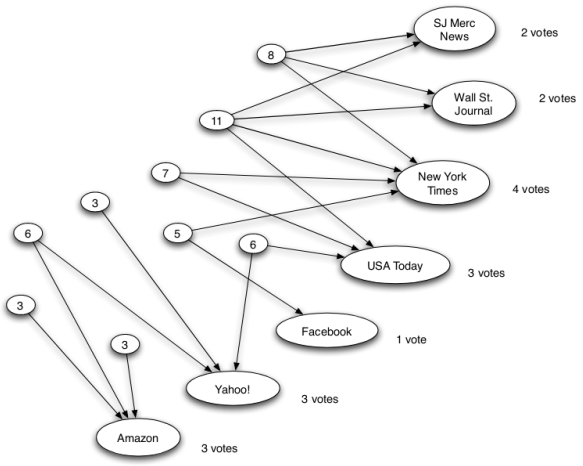



Figura 61: Trovare buone liste per la ricerca "quotidiani".

### 14.2.3 The Principle or Repeated Improvement

Se pensiamo che le pagine lista siano importanti, dobbiamo allora aumentare il peso del loro voto, usando come valore il valore della lista. La figura 62 mostra cosa succede seguendo questo metodo: i principali quotidiani ora superano le liste invece di essere soppressi da esse.

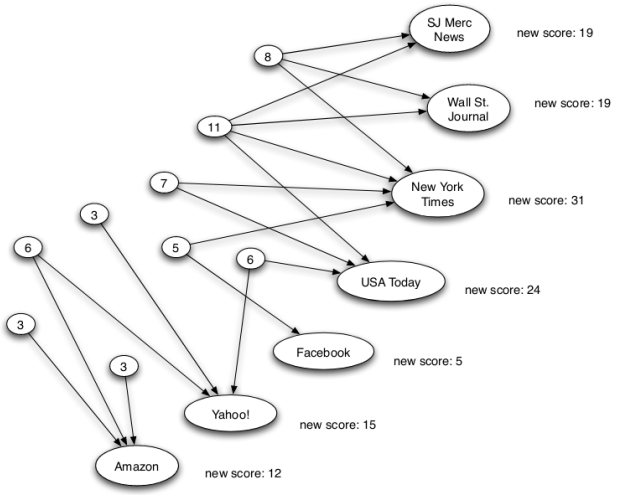



Figura 62: Il nuovo calcolo del valore delle pagine usando come peso delle liste il loro valore.

Adesso che abbiamo voti migliori per le pagine richieste, possiamo usarli per migliorare il voto delle liste, e viceversa, in un ciclo chiamato **principle**

**of repeated improvement**.

### 14.2.4 Hubs and Authorities

Possiamo formalizzare questa procedura, con il nome di **hits**, come segue: chiameremo **authorities** le pagine che si stavano effettivamente cercando tramite la ricerca, e **hubs** la pagine lista che puntano ad esse. Per ogni pagina $p$ cerchiamo di stimare entrambi questi valori, chiamati *auth(p)* e *hub(p)*, partendo da 1, ovvero indifferenza. Iniziamo quindi la fase di voto in cui usiamo gli hubs per stimare le authorities secondo questa regola:

> Authority Update Rule: per ogni pagina $p$, aggiorna *auth(p)* sommando i valori degli hubs che puntano ad essa.

La regola invece per stimare il valore degli hubs usando le authorities è la seguente:

> Hub Update Rule: per ogni pagina $p$, aggiorna *hub(p)* sommando i valori delle authorities che puntano ad essa.

Ripetiamo questo procedimento per $k$ volte, ottenendo però dei valori molto alti. Per questo **normalizziamo** i dati ottenuti dividendo il valore di ogni authority per la somma dei valori di tutte le authorities, e ripetendo il procedimento per ogni hub.

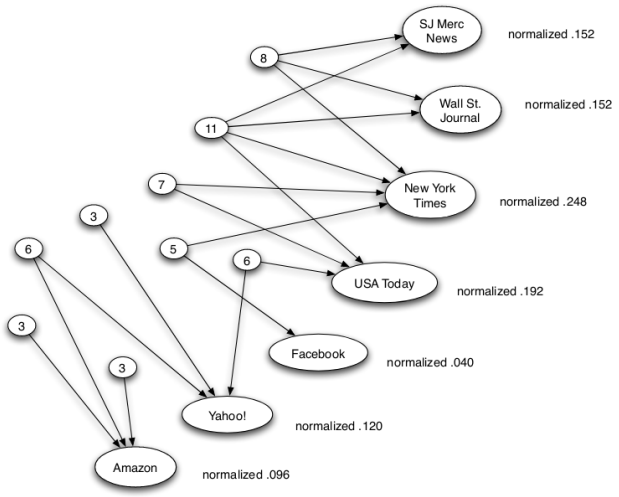



Figura 63: I valori di classificazione normalizzati.

Portando all'estremo questo procedimento, scegliendo grandi valori di $k$, i valori convergono al loro limite, rendendo i cambiamenti di ogni iterazione po-

co significativi. Questi valori limite corrispondono ad una sorta di equilibrio tra i valori di hubs e authorities, proporzionali tra loro.

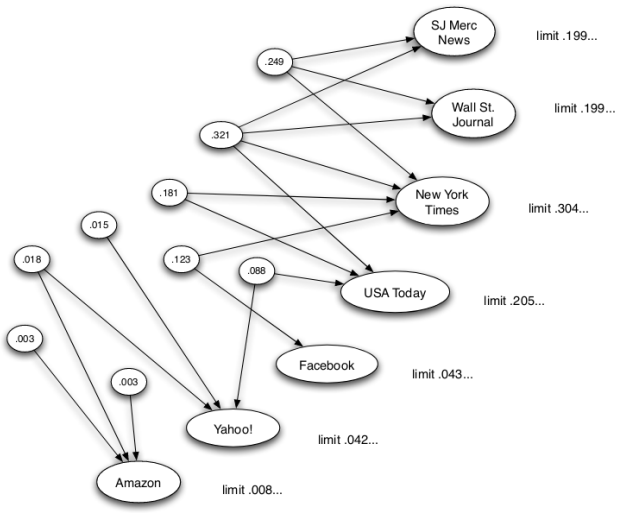



Figura 64: Il limite dei valori di classificazione.

## 14.3 Page Rank

L'idea dietro hubs e authorities è quindi che le pagine possono ricoprire ruoli diversi, ed appoggiare altre pagine senza essere a loro volta appoggiate. Questo è il caso di diverse authorities concorrenti, che hanno interesse a prevalere sulle altre, e sono quindi raggruppabili solo grazie agli hubs che puntano ad esse.

In altri contesti nel Web l'approvazione tra pagine è invece vista più come il passaggio tra una pagina importante ad un'altra, ovvero una pagina è importante se è citata da altre pagine importanti. Questo è il caso di pagine governative e accademiche o di blog personali, ed è alla base della misura di importanza per il **PageRank**.

### 14.3.1 The Basic Definition of PageRank

Possiamo pensare al PageRank come una sorta di "fluido" che circola nella rete, attraversando i diversi nodi e raccogliendosi in quelli principali. Esso è calcolato nel modo seguente:

- in una rete con $n$ nodi, assegnamo a tutti lo stesso PageRank iniziale, pari a $1/n$;

- scegliamo un numero $k$ di passi;
- effettuiamo una serie di $k$ aggiornamenti ai valori di PageRank, usando la seguente regola:

  > Basic PageRank Update Rule: ogni pagina divide il proprio PageRank equamente tra i suoi collegamenti uscenti e ne passa una parte ad ognuno. Ogni pagina aggiorna il proprio PageRank alla somma delle parti che ottiene.

Si noti che alla fine di questo processo il PageRank totale della rete sarà il medesimo, poiché esso non viene creato né distrutto. Vediamo un esempio:

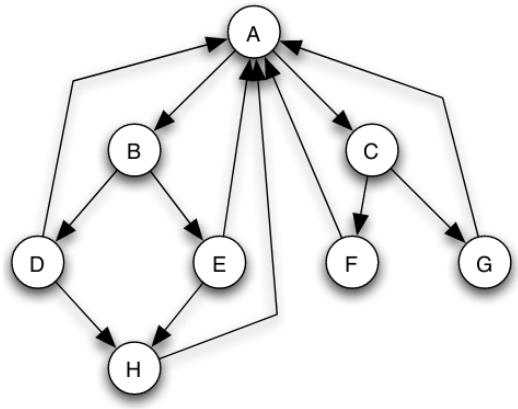



Figura 65: Un insieme di esempio per il calcolo del PageRank.

| Step | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | 1/2 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/8 |
| 2 | 3/16 | 1/4 | 1/4 | 1/32 | 1/32 | 1/32 | 1/32 | 1/16 |

Figura 66: I primi due round di aggiornamento del PageRank.

### 14.3.2 Equilibrium Values of PageRank

Come nel caso di hubs e authorities, anche il valore di PageRank di ogni nodo converge ad un certo limite per un numero grande di $k$. Poiché il PageRank è conservato in ogni computazione, con somma totale di tutti i valori pari ad 1, possiamo pensare che il limite di ogni nodo porti ad una sorta di **equilibrio**. Siamo in questo stato se la somma di tutti i PageRank è 1 e applicando un nuovo aggiornamento i valori non cambiano.

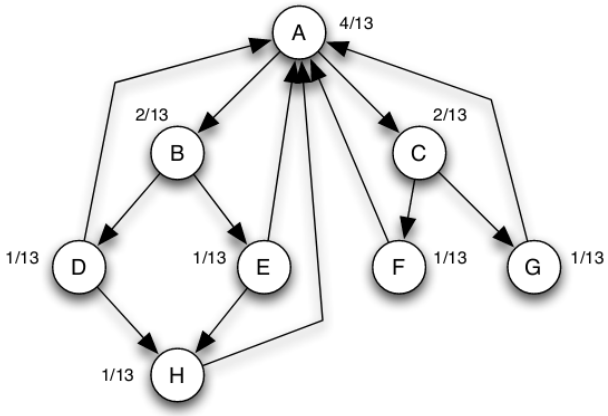



Figura 67: L'equilibrio dei valori di PageRank.

### 14.3.3 Scaling the Definition of PageRank

Spesso però può succedere che i nodi sbagliati abbiano un alto valore di PageRank, ma esiste un modo semplice per risolvere questo problema. Se ad esempio prendiamo la rete nella figura 65 e puntiamo F e G verso l'un l'altro invece che verso A, notiamo che si forma un pozzo, ed il PageRank che arriva a quei due nodi non viene mai redistribuito verso la rete. Serve quindi uno **scaling factor** $s$ compreso tra 0 ed 1 per modificare la precedente definizione di PageRank come segue:

> Scaled PageRank Update Rule: prima si applica la regola di PageRank base. Poi si riducono proporzionalmente ad una variabile $s$ tutti i valori di PageRank, riducendo la somma di tutti i valori da 1 ad $s$. Si distribuiscono poi i valori di PageRank rimanenti, cioè (1 - $s$) a tutti i nodi nella rete in maniera eguale.

Questa procedura permette di evitare che si formino dei pozzi di PageRank, ridistribuendone una parte in tutta la rete senza perdere il valore totale.

### 14.3.4 Random Walks: An Equivalent Definition of PageRank

Vediamo ora una descrizione equivalente di PageRank, ma diversa in termini di formulazione. Consideriamo una navigazione nel Web, iniziando da una pagina casuale. Da questa si seguono dei link casuali per $k$ pagine, arrivando ogni volta in una nuova pagina. Questo tipo di esplorazione è chiamata **random walk** nella rete. Dimostreremo che:

> Claim: la probabilità di essere alla pagina $X$ dopo $k$ passi di una random walk è precisamente il valore di PageRank di $X$ dopo $k$ passi della regola base di aggiornamento del PageRank.

Come già detto le due definizioni sono equivalenti, ma spesso vedere il problema da una nuova prospettiva porta a nuove idee ed intuizioni sui diversi problemi, come quello del pozzo di cui abbiamo parlato prima.

## 14.4 Applying Link Analysis in Modern Web Search

Le procedure di analisi dei collegamenti descritte in precedenza erano ottime negli anni '90, ma con la crescita esponenziale del Web negli ultimi anni è servito espandere e generalizzare questo processo. È difficile parlare degli attuali algoritmi di analisi, sia perché essi sono in continua evoluzione, sia perché le compagnie dei motori di ricerca non rendono pubblichi i loro dati. In particolare il PageRank è sempre stato alla base della metodologia di Google, venendo poi affiancata da altre metodologie complementari, tra cui si pensa ci sia quella chiamata "Hilltop", un'estensione di hubs e authorities.

### 14.4.1 Combining Links, Text, and Usage Data

Per ottenere dei buoni risultati è quindi importante combinare i collegamenti delle pagine ed il loro contenuto testuale. Un modo per fare ciò è l'analisi dell'**anchor text**, la parte evidenziata di un testo cliccabile che attiva il collegamento ad un'altra pagina.

Il metodo descritto in precedenza per il calcolo del PageRank tramite hubs e authorities può essere esteso per comprendere questi nuovi dati, semplicemente pesando il voto dei collegamenti con degli importarti anchor texts, ad esempio moltiplicando i valori di PageRank passati di un fattore che ne indica la qualità.

### 14.4.2 A Moving Target

Se pensiamo ad esempio al fatto che i risultati economici e di notorietà di una azienda possono dipendere molto dall'ordine dei risultati delle ricerche degli utenti, è normale pensare che, nell'ottica di rimanere in vetta o incrementare la loro notorietà, le aziende cerchino sempre di modificare l'authoring style delle loro pagine per andare incontro al sistema di rank dei motori di ricerca.

Nonostante inizialmente questa pratica non fosse ben accetta, col tempo è diventata regolarizzata e standardizzata, portando alla nascita di linee guida per ottenere migliori risultati. Nacque anche una grande industria nota

come **search engine optimization** (SEO), formata da diversi esperti che guidavano le compagnie a creare pagine e siti di un alto rank.

Questo processo ebbe però diverse conseguenze: per i motori di ricerca la "perfetta" funzione di ranking rimarrà sempre variabile; inoltre essi terranno segreta la loro funzione di ranking, sia per questioni di competizione con altri motori, sia per non promuovere la ricerca di un rank migliore da parte degli interessati. Inoltre la crescita di questo settore ha portato ad un modello di business basato sulla pubblicità: molti motori di ricerca offrono la possibilità di essere inseriti tra i migliori risultati in cambio di un pagamento.

## 14.5 Applications beyond the Web

Le tecniche di analisi dei collegamenti viste in questo capitolo vengono usate anche in diversi altri contesti, a patto che si parli di una struttura a forma di rete.

### 14.5.1 Citation Analysis

Come già visto in precedenza, lo studio delle citazioni nei papers di ricerca ha un'origine molto più recente del Web. Una misura standard in questo ambito è l'**impact factor** di Garfield, definito come il numero medio di citazioni ricevute da un paper da parte di una rivista scientifica negli ultimi due anni.

Negli anni '70 questa definizione fu rivisitata ed estesa, partendo dall'idea che non tutti i voti debbano avere lo stesso peso, ma che le citazioni da giornali più importanti debbano valere di più. Questo concetto fu formulato come **influence weights** per le riviste, ed è molto simile al concetto di PageRank per il Web.

### 14.5.2 Link Analysis of U.S. Supreme Court Citations

Di recente i ricercatori hanno applicato la tecnica di analisi dei collegamenti anche alla rete di citazioni tra le decisioni legali dei tribunali americani, un ambito in cui le citazioni sono fondamentali per capire e motivare le decisioni prese. Inoltre i tribunali offrono spunti per osservare come le decisioni prese cambiano durante il tempo.

## 14.6 Advanced Material: Spectral Analysis, Random Walks, and Web Search

In questa sezione vengono analizzati i metodi per calcolare i valori di hub, authority e PageRank. L'utilizzo di autovalori ed autovettori per studiare la struttura delle reti è noto come **spectral analysis** dei grafi.

### 14.6.1 Spectral Analysis of Hubs and Authorities

Dimostriamo come i valori di hubs e authorities convergono al loro limite dopo un certo numero di computazioni.

**Adjacency Matrices and Hub/Authority Vectors.** Consideriamo un insieme di $n$ pagine, considerate come nodi, e rappresentiamo in una matrice $M$ di dimensione $n$x$n$ gli archi tra essi. Questa matrice, chiamata **adjacency matrix**, contiene come elemento $M_{i,j}$ il valore 1 se vi è un arco tra il nodo $i$ ed il nodo $j$, 0 in caso contrario. Con due vettori $a$ e $b$ invece rappresentiamo i valori di hubs e authorities per ogni nodo.

Vediamo un esempio nella figura 68.

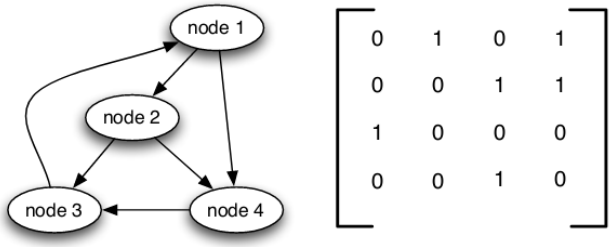



Figura 68: Un esempio di insieme di pagine Web con la rispettiva matrice di adiacenza.

**Hub and Authority Update Rules as Matrix-Vector Multiplication.** Consideriamo la regola di aggiornamento degli hubs usando la notazione appena definita. Per ogni nodo $i$, il suo valore di hub $h_i$ viene aggiornato come la somma di $a_j$:

$$h_i \leftarrow M_{i1}a_1 + M_{i2}a_2 + \ldots + M_{in}a_n$$

che corrisponde ad una moltiplicazione matrice per vettore, ovvero:

$$\mathrm{h} \leftarrow \mathrm{Ma}$$

La figura 69 mostra un esempio di questo passo.

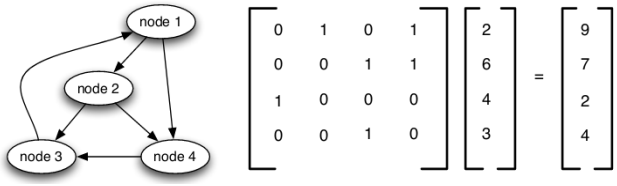



Figura 69: L'aggiornamento degli hubs.

Allo stesso modo possiamo aggiornare le authorities, in modo che $a_j$ sia la somma dei valori di tutti gli hubs collegati:

$$a_i \leftarrow M_{1i}h_1 + M_{2i}h_2 + \ldots + M_{ni}h_n$$

che in realtà corrisponde alla moltiplicazione tra la matrice trasposta ed il vettore h:

$$\mathrm{h} \leftarrow \mathrm{M}^T\mathrm{a}$$

**Unwinding the k-Step Hub-Authority Computation.** Cosa succede se però effettuiamo $k$ passi dell'algoritmo, per grandi valori di $k$? Denotiamo con $\mathrm{a}^{\langle 0\rangle}$ e $\mathrm{h}^{\langle 0\rangle}$ i vettori iniziali con tutti i valori pari ad 1, e con $\mathrm{a}^{\langle k\rangle}$ e $\mathrm{h}^{\langle k\rangle}$ i vettori dopo $k$ applicazioni della regola. Otteniamo quindi la seguente sequenza:

$$a^{\langle 1\rangle} = M^T h^{\langle 0\rangle}$$
$$h^{\langle 1\rangle} = Ma^{\langle 1\rangle} = MM^T h^{\langle 0\rangle}$$

come primo passo, e continuando:

$$a^{\langle 2\rangle} = M^T h^{\langle 1\rangle} = M^T MM^T h^{\langle 0\rangle}$$
$$h^{\langle 2\rangle} = Ma^{\langle 2\rangle} = MM^T MM^T h^{\langle 0\rangle} = (MM^T)^2 h^{\langle 0\rangle}$$
$$a^{\langle 3\rangle} = M^T h^{\langle 2\rangle} = M^T MM^T MM^T h^{\langle 0\rangle} = (M^T M)^2 M^T h^{\langle 0\rangle}$$
$$h^{\langle 3\rangle} = Ma^{\langle 3\rangle} = MM^T MM^T MM^T h^{\langle 0\rangle} = (MM^T)^3 h^{\langle 0\rangle}$$

e così via, arrivando alla regola generale:

$$a^{\langle k\rangle} = (M^T M)^{k-1} M^T h^{\langle 0\rangle}$$
$$h^{\langle k\rangle} = (MM^T)^k h^{\langle 0\rangle}$$

**Thinking About Multiplication in Terms of Eigenvectors.** Perché un valore di hub o authority converga, serve normalizzare i dati. Ciò equivale a dire che è la **direzione** dei vettori $h$ e $a$ che converge. Ci sono quindi due costanti $c$ e $d$ tali che $\frac{h^{\langle k \rangle}}{c^k}$ e $\frac{a^{\langle k \rangle}}{d^k}$ convergono ai limiti per $k$ che va ad infinito.

Considerando i vettori di hub, abbiamo che se:

$$\frac{h^{\langle k \rangle}}{c^k} = \frac{(MM^T)^k h^{\langle 0 \rangle}}{c^k}$$

convergesse al limite $h^{\langle * \rangle}$, ci dovremmo aspettare che la direzione di $h^{\langle * \rangle}$ non cambi se moltiplicata per $(MM^T)$, mentre la lunghezza aumenterà di un fattore $c$. Perché questo valga, $h^{\langle * \rangle}$ deve soddisfare l'equazione:

$$(MM^T)h^{\langle * \rangle} = ch^{\langle * \rangle}$$

Il vettore che soddisfa questa equazione è chiamato **autovettore**, mentre il fattore $c$ è chiamato **autovalore**.

Per dimostrare questa proprietà useremo la seguente nozione:

Ogni matrice simmetrica $A$ con $n$ righe ed $n$ colonne ha un insieme di $n$ autovettori unitari e tutti mutualmente ortogonali, ovvero formanti una **base** per lo spazio $\mathbf{R}^n$.

Dato che $MM^T$ è una matrice simmetrica, ossia una matrice $A$ in cui $A_{i,j} = A_{j,i}$, possiamo applicare la proprietà. Denotiamo con $z_1, z_2, \ldots, z_n$ gli autovettori mutualmente ortogonali e con $c_1, c_2, \ldots, c_n$ gli autovalori corrispondenti, in modo che $|c_1| \geq |c_2| \geq \ldots \geq |c_n|$ e per semplicità che $|c_1| > |c_2|$. Ora, dato qualsiasi vettore $x$, per fare il prodotto matrice - vettore conviene esprimere $x$ come combinazione lineare dei vettori $z$, ovvero $x = p_1 z_1 + p_2 z_2 + \ldots + p_n z_n$, in modo da ottenere:

$$\begin{aligned}(MM^T)x &= (MM^T)(p_1 z_1 + p_2 z_2 + ... + p_n z_n) \\ &= p_1 MM^T z_1 + p_2 MM^T z_2 + \ldots + p_n MM^T z_n \\ &= p_1 c_1 z_1 + p_2 c_2 z_2 + \ldots + p_n c_n z_n\end{aligned}$$

Ciò permette di analizzare più facilmente la moltiplicazione di $MM^T$ per grandi potenze.

**Convergence of the Hub-Authority Computation.** Quando moltiplichiamo più volte il vettore $x$ per $MM^T$ otteniamo un risultato di questo tipo:

$$(MM^T)^k x = c_1^k p_1 z_1 + c_2^k p_2 z_2 + \ldots + c_n^k p_n z_n$$

Pensando al contesto degli hubs, ricordiamo che $h^{\langle k \rangle} = (MM^T)^k h^{\langle 0 \rangle}$ , dove $h^{\langle 0 \rangle}$ è il vettore iniziale pari ad 1, che può essere rappresentato in termini di vettori di base $z_1, z_2, \ldots, z_n$ come combinazione lineare:

$$h^{\langle 0 \rangle} = q_1 z_1, q_2 z_2, \ldots, q_n z_n$$

Quindi:

$$h^{\langle k \rangle} = (MM^T)^k h^{\langle 0 \rangle} = c_1^k q_1 z_1 + c_2^k q_2 z_2 + \ldots + c_n^k q_n z_n$$

e dividendo entrambi i lati per $c_1^k$ otteniamo:

$$\frac{h^{\langle k \rangle}}{c_1^k} = q_1 z_1 + (\frac{c_2}{c_1})^k q_2 z_2 + \ldots + (\frac{c_n}{c_1})^k q_n z_n$$

Dato che $|c_1| > |c_2|$, al crescere di $k$ verso infinito, ogni termine eccetto il primo vale 0. In conseguenza, per $k \to \infty$, la sequenza $\frac{h^{\langle k \rangle}}{c_1^k}$ vale $q_1 z_1$.

**Wrapping Up.** Per arrivare alla convergenza dobbiamo però fare alcune considerazioni. La prima è che il risultato è indipendente dal vettore iniziale $h^{\langle 0 \rangle}$ scelto, ossia che la direzione del risultato non cambia in base alla direzione del vettore iniziale. Supponiamo che il vettore scelto $x$ sia positivo $x = p_1 z_1 + p_2 z_2 + \ldots + p_n z_n$, quindi $(MM^T)x = c_1^k p_1 z_1 + c_2^k p_2 z_2 + \ldots + c_n^k p_n z_n$ e $\frac{h^{\langle k \rangle}}{c_1^k}$ converge a $p_1 z_1$, un vettore di direzione $z_1$, anche con la scelta di un vettore iniziale diverso.

La seconda considerazione è che $p_1$ e $q_1$ non siano nulli; se prendiamo $x = p_1 z_1 + p_2 z_2 + \ldots + p_n z_n$ e facciamo il prodotto interno tra $z_1$ e $x$, considerando che i vettori $z$ sono ortogonali tra loro, otteniamo:

$$z_1 \cdot x = z_1(p_1 z_1 + p_2 z_2 + \ldots + p_n z_n) = p_1(z_1 \cdot z_1 + p_2(z_1 \cdot z_2) + \ldots + p_n(z_1 \cdot z_n) = p_1$$

poiché tutti i termini della somma sono 0 tranne il primo che vale $p_1$, che è il prodotto interno tra $x$ e $z_1$. Questo mostra che la sequenza di vettori di hub converge a un vettore non nullo nella direzione di $z_1$, ammesso che il vettore iniziale non sia ortogonale a $z_1$ stesso.

Vediamo però ora che nessun vettore positivo può essere ortogonale a $z_1$ tramite i seguenti passi:

1. non è possibile per ogni vettore essere ortogonale a $z_1$, per cui esiste qualche vettore $x$ positivo tale che $(MM^T)^k \frac{x}{c_1}$ converga a un vettore non nullo $p_1 z_1$;

2. dato che $(MM^T)^k \frac{x}{c_1}$ comprende solo numeri non negativi, $p_1 z_1$ deve avere solo coordinate non negative ed almeno una positiva perché non è uguale a 0;

3. se consideriamo quindi il prodotto interno di ogni vettore positivo con $p_1 z_1$, il risultato dev'essere positivo. Di conseguenza nessun vettore positivo può essere ortogonale a $z_1$. Ciò stabilisce che la sequenza di vettori di hub converge al vettore in direzione di $z_1$ quando partiamo da un qualsiasi vettore positivo.

Rimane ora da allentare il vincolo $|c_1| > |c_2|$. In generale possono esserci $l > 1$ autovalori legati al più grande valore assoluto: possiamo avere $|c_1| = \ldots = |c_l|$ e gli autovalori $c_{l+1}, \ldots, c_n$ tutti minori in valore assoluto. Abbiamo quindi che $c_1 = \ldots = c_l > c_{l+1} \geq \ldots \geq c_n \geq 0$, allora:

$$\frac{h^{\langle k \rangle}}{c_1^k} = q_1 z_1 + \ldots + q_l z_l + (\tfrac{c_{l+1}}{c_1})^k q_{l+1} z_{l+1} + \ldots + (\tfrac{c_n}{c_1})^k q_n z_n$$

La somma dei termini da $l + 1$ ad $n$ vale 0, quindi la sequenza converge alla somma dei termini $q_1 z_1 + \ldots + q_l z_l$. Quindi quando $c_1 = c_2$ abbiamo la convergenza ma il limite può dipende dalla scelta del vettore iniziale. Nella realtà però per strutture di collegamenti sufficientemente grandi si ha quasi sempre una matrice $M$ con la proprietà che $MM^T$ abbia $|c_1| > |c_2|$.

Tutto questo discorso può anche essere adattato alle authorities, considerando però il prodotto $M^T M$.

### 14.6.2 Spectral Analysis of PageRank

Vediamo adesso come analizzare il PageRank usando un procedimento simile. Il flow del PageRank, che ogni nodo divide equamente ed invia ai nodi adiacenti, può essere rappresentato con una matrice $N$, ed $N_{i,j}$ rappresenta il flow di $i$ che deve essere inviato a $j$ in un passo. In altre parole $N_{i,j}$ è uguale a $\frac{1}{l_i}$, dove $l_i$ rappresenta il numero di connessioni del nodo $i$; col vettore $r$ rappresentiamo invece il PageRank di ogni nodo. In questo modo abbiamo:

$$r_i \leftarrow N_{1i} r_1 + N_{2i} r_2 + \ldots + N_{ni} r_n$$

che corrisponde alla moltiplicazione per il trasposto della matrice, quindi:

$$r \leftarrow N^T r$$

Indichiamo invece con $\tilde{\mathrm{N}}$ la matrice usata per la scaled update rule, avendo così $\tilde{\mathrm{N}}_{ij} = sN_{ij} + (1 - s)/n$, e come regola:

$$r_i \leftarrow \tilde{\mathrm{N}}_{1i} r_1 + \tilde{\mathrm{N}}_{2i} r_2 + \ldots + \tilde{\mathrm{N}}_{ni} r_n$$

oppure:

$$r \leftarrow \tilde{\mathrm{N}}^T r$$

**Repeated Improvement Using the Scaled PageRank Update Rule.** Applicando più volte la regola, otteniamo che:

$$r^{\langle k \rangle} = (\tilde{\mathrm{N}}^T)^k r^{\langle 0 \rangle}$$

Dato che il valore totale del PageRank non cambia, non serve normalizzare i vettori.

Analogamente ad hubs e authorities, ci aspettiamo che la regola di aggiornamento converga al vettore limite $r^{\langle * \rangle}$, tale che $\tilde{\mathrm{N}}^T r^{\langle * \rangle} = r^{\langle * \rangle}$, ossia che $r^{\langle * \rangle}$ sia un autovettore di $\tilde{\mathrm{N}}^T$ con autovalore corrispondente pari ad 1.

**Convergence of the Scaled PageRank Update Rule.** Per le matrici come $\tilde{\mathrm{N}}$ con tutti gli elementi positivi possiamo usare il **teorema di Perron**, che afferma che ogni matrice $P$ con tutti gli elementi positivi ha le seguenti proprietà:

1. $P$ ha un autovalore reale $c > 0$ tale che $c > |c'|$ per tutti gli altri autovalori $c'$;

2. esiste un autovettore $y$ con coordinate reali positive corrispondenti al più grande autovalore $c$, ed esso è unico fino alla moltiplicazione per una costante;

3. se il più grande autovalore $c$ è uguale ad 1, per ogni vettore iniziale $x \neq 0$ con coordinate non negative, la sequenza di vettori $P^k x$ converge ad un vettore della direzione di $y$ per $k$ che va ad infinito.

Questo teorema ci dice quindi che esiste un unico vettore $y$ che rimane fisso per le diverse applicazioni della regola di aggiornamento, e l'applicazione della regola su un qualsiasi vettore iniziale converge ad $y$, il vettore che contiene i limiti dei valori di PageRank.

### 14.6.3 Formulation of PageRank Using Random Walks

Ora ci occupiamo di formulare il PageRank in termini di random walk. Se $b_1, b_2, \ldots, b_n$ denotano rispettivamente la probabilità di essere al nodo 1, 2, $\ldots$, $n$ ad un determinato passo, qual è la probabilità di essere al nodo $i$ al passo successivo? Possiamo rispondere così:

1. per ogni nodo $j$ che porta ad $i$, che una probabilità di $1/l_j$ di muoversi verso quest'ultimo;

2. perché questo succeda bisogna effettivamente essere al nodo $j$, quindi questo contribuisce con una probabilità $b_j(1/l_j) = b_j/l_j$ di essere ad $i$ nel passo successivo;

3. sommando $b_j/l_j$ su tutti i nodi $j$ che portano ad $i$ si ottiene la probabilità di essere a $b_i$ nel passo successivo.

Possiamo usare la matrice $N$ usata in precedenza per definire la regola di aggiornamento:

$$b_i \leftarrow N_{1i}b_1 + N_{2i}b_2 + \ldots + N_{ni}b_n$$

oppure come moltiplicazione matrice vettore:

$$r \leftarrow N^T b$$

Scopriamo che si tratta della stessa regola di aggiornamento base vista in precedenza, dato che si parte dagli stessi valori di PageRank e di probabilità. Possiamo quindi affermare:

Claim: la probabilità di essere alla pagina $X$ dopo $k$ passi di random walk è esattamente il PageRank di $X$ dopo $k$ applicazioni della regola di aggiornamento base.

In altre parole, in entrambi i casi il flow del PageRank si muove nello stesso modo.

**A Scaled Version of the Random Walk.** Possiamo interpretare anche la scaled update rule in termini di random walk, ammettendo che per un valore $s > 0$, con probabilità $s$ il cammino sceglie un nodo casuale come in precedenza, mentre con probabilità (1 - $s$) si salta ad un nodo casuale. Come in precedenza, $b_1, b_2, \ldots, b_n$ denotano rispettivamente la probabilità di essere al nodo 1, 2, ..., $n$ ad un determinato passo. La probabilità di essere al nodo $i$ in un determinato passo è la somma di $sb_j/l_j$ su tutti i nodi $j$ che portano ad $i$, più (1 - $s$)/$n$. Usando la matrice $\tilde{\mathrm{N}}$ come in precedenza otteniamo:

$$b_i \leftarrow \tilde{\mathrm{N}}_{1i}b_1 + \tilde{\mathrm{N}}_{2i}b_2 + \ldots + \tilde{\mathrm{N}}_{ni}b_n$$

oppure:

$$b \leftarrow \tilde{\mathrm{N}}^T b$$

Anche in questo caso la regola di aggiornamento è la medesima perché si usano gli stessi valori iniziali e la stessa regola di aggiornamento:

Claim: la probabilità di essere alla pagina $X$ dopo $k$ passi di scaled random walk è esattamente il PageRank di $X$ dopo $k$ applicazioni della scaled update rule.

Quest afferma anche che per un numero di passi che va ad infinito, il limite della probabilità di essere al nodo $X$ è uguale al limite del valore di PageRank di $X$.

# 18 Power Laws and Rich-Get-Richer Phenomena

## 18.1 Popularity as a Network Phenomenon

Quando il comportamento in un sistema è correlato alla popolazione può portare a risultati diversi rispetto a quelli di sistemi in cui un individuo sceglie solo per se stesso. Con questo concetto analizziamo la nozione di **popolarity**, un fenomeno che interessa una parte ristretta della popolazione che ottiene un'ampia visibilità, ed una parte ancora più ristretta che è conosciuta ovunque.

Un contesto in cui la popolarità di può facilmente osservare e misurare è il Web, dove il numero di in-links può essere un valido fattore di popolarità. Il quesito da porsi è:

Come funzione di $k$, quale frazione delle pagine nel Web ha $k$ in-links?

### 18.1.1 A Simple Hypothesis: The Normal Distribution

Normalmente dovremmo aspettarci una distribuzione *normale* oppure *Gaussiana*. La figura 70 mostra una distribuzione normale con media pari a 0 e deviazione standard pari ad 1. La particolarità di questa distribuzione è che la probabilità di osservare un che supera la media di più di $c$ volte deviazione standard decresce esponenzialmente in $c$.

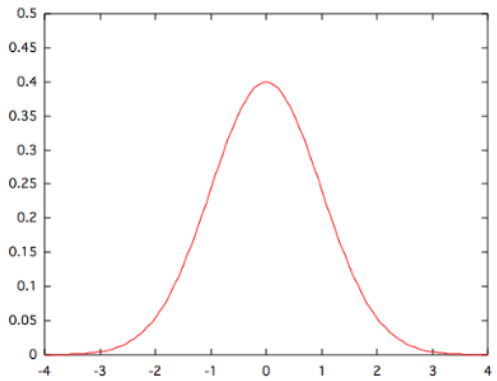



Figura 70: La densità dei valori in una distribuzione normale.

Il Central Limite Theorem, formulato agli inizi del 1900, spiega perché una distribuzione simile sia presente in così tanti diversi contesti. Esso afferma

che se si prende una sequenza di piccole quantità indipendenti e le si somma, al limite la loro somma è distribuita secondo una distribuzione normale.

Possiamo applicare questo teorema considerando che ogni pagina del Web decida casualmente di collegarsi ad un'altra pagina. In questo modo la quantità di in-links di una pagine è la somma di diverse quantità indipendenti e dovremmo aspettarci una distribuzione normale, quindi il numero di pagine con $k$ in-links dovrebbe decrescere esponenzialmente in $k$, al crescere di $k$.

## 18.2 Power Laws

Misurando però l'effettiva distribuzione dei collegamenti nel Web i risultati sono diversi da quelli predetti con in central limit theorem. Ciò che si ottiene è che la frazione di pagine con $k$ in-links è proporzionale a circa $1/k^2$ (l'esponente è in genere compreso tra 2 e 3). Questo perché $1/k^2$ diminuisce molto più lentamente al crescere di $k$, quindi le pagine con grandi quantità di in-links sono più comuni di quanto ci si aspetti. Una funzione che diminuisce come $k$ ad una potenza fissata è chiamata **power law**.

La distribuzione della popolarità è quindi chiaramente sbilanciata, e questo porta al nascere di valori molto alti; inoltre secondo il ragionamento precedente sul Web, c'è sicuramente un grande numero di pagine molto popolari.

Le power laws sono molto comuni anche il altri settori, soprattutto nei casi in cui la quantità studiata può essere vista come una popolazione. C'è una prova molto semplice da effettuare per vedere se un insieme di dati comprende una distribuzione power-law. Sia $f(k)$ la frazione di oggetti con valore $k$ e supponiamo di voler sapere se valga l'equazione $f(k) = a/k^c$, per qualche esponente $c$ e costante di proporzionalità $a$. Riscrivendo come $f(k) = ak^{-c}$ e usando il logaritmo in entrambi i lati otteniamo:

$$\log f(k) = \log a - c \log k$$

Questo ci dice se abbiamo una power law e tracciando $\log f(k)$ come una funzione di $\log k$, dovremmo vedere una linea dritta: -$c$ indica la pendenza e $\log a$ l'intersezione con l'asse $y$. Vediamo un esempio nella figura 71.

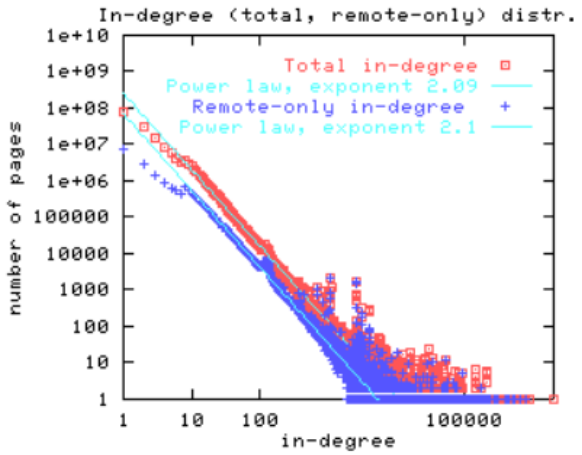



Figura 71: Una distribuzione power-law che mostra una retta in un grafico log-log.

Se però ammettiamo che la distribuzione power-law sia così diffusa, dobbiamo cercare di capire il motivo che sta alla base di questo fatto.

## 18.3 Rich-Get-Richer Models

Come la distribuzione normale nasce dalle scelte casuali, le power laws nascono dal processo di decisione delle persone. Fornire un modello preciso per questo ambito è difficile, perciò ci baseremo sul concetto che le persone tendono a ripetere le decisioni prese da altre persone prima di loro.

Partendo da queste idee, si può sviluppare un modello per la creazione di collegamenti tra pagine Web in questo modo:

1. le pagine vengono create in ordine e numerate da 1 ad $N$;
2. quando la pagina $j$ viene creata, produce un collegamento con una pagina create in precedenza scegliendo un'azione tra (a) e (b) riportate in seguito secondo la seguente regola probabilistica:
   (a) con probabilità $p$, la pagina $j$ sceglie la pagina $i$ casualmente tra tutte le pagine già esistenti;
   (b) con probabilità 1 - $p$, la pagina $j$ sceglie casualmente un pagina $i$ e crea un collegamento alla pagina a cui $i$ punta;
   (c) questo descrive la creazioen della pagina $j$; ripetendo questo processo si possono creare collegamenti multipli ed indipendenti generati dalla pagina $j$.

Il risultato di questo modello è che usandolo per molte pagine, si ottiene una distribuzione power-law. Notiamo che la parte 2(b) esprime la volontà di copiare quanto fatto da un nodo precedente, e questo porta alla dinamica **richer-get-richer**, che esprime il fatto che la probabilità che una pagina $l$ incrementi la sua popolarità è direttamente proporzionale alla popolarità attuale di $l$. Questo fenomeno è anche noto col nome di **preferential attachment**. La differenza col central limit theorem è che valori piccoli indipendenti e casuali tendono ad annullarsi, mentre col fenomeno richer-get-richer si tende ad ampliare ancora di più i valori già grandi in precedenza, e questo porta ad una distribuzione power-law.

Questo concetto si può estendere anche a concetti che non hanno nulla a che fare con le decisioni umane, come ad esempio la popolazione di una città, il numero di geni in un genoma o l'esempio visto in precedenza delle pagine Web.

## 18.4 The Unpredictability of Rich-Get-Richer Effects

Data la natura degli effetti che producono le power laws è facile pensare che la situazione iniziale del sistema sia relativamente fragile, mentre quando questi aspetti sono consolidati il processo richer-get-richer produce i suoi effetti. Bisogna però pensare che la situazione casuale iniziale abbia quindi un ruolo importante nella formazione di questi sistemi. Secondo questa intuizione, se si potesse tornare indietro nel tempo e rivivere la storia, la distribuzione di popolarità sarebbe la stessa, ma i fenomeni interessati sarebbero diversi da ora.

Sebbene sia difficile simulare questi aspetti, di recente Salgankik, Dodds e Watts fecero un esperimento creando un sito di distribuzione musicale con 48 canzoni sconosciute create da gruppi veri. Ogni utente del sito aveva la possibilità di ascoltare le canzoni e poteva vedere quante volte ogni brano fosse stato scaricato. All'arrivo sulla piattaforma però ogni utente era assegnato ad una di otto copie parallele del sito, ognuna delle quali cominciava nello stesso modo ed evolveva diversamente. Questo ha permesso di mostrare come la popolarità delle canzoni potesse variare attraverso le diverse copie parallele. Una nona copia, in cui il numero di download non era mostrato, ha invece dimostrato che in questo ambiente la variazione di popolarità non era così accentuata.

### 18.4.1 Closer Relationships Between Power Laws and Information Cascades?

Non abbiamo ancora trattato le information cascades, ma il concetto si può riassumere come l'influenza che le scelte precedenti tra due opzioni hanno sulla scelta attuale di un determinato individuo. Anche se questo sceglie l'opzione ottima, si può arrivare ad un effetto a cascata su una delle due opzioni. Per legare le information cascades con le power laws però si dovrebbe offrire la possibilità di scegliere tra molti più di due elementi, e bisognerebbe guardare alla decisione di una sola persona e non tutte le precedenti. Riuscire a superare queste differenze potrebbe offrire nuovi spunti ed informazioni sulla crescita della popolarità in base alla dinamica richer-get-richer e la distribuzione power-law.

## 18.5 The Long Tail

La distribuzione di popolarità può avere importanti conseguenze sul piano economico. Si pensi ad esempio al business di un'azienda che vende libri: la maggior parte delle vendite deriva da pochi oggetti molto popolari, oppure da tutto l'insieme restante degli oggetti meno popolari? Nel 2004 un articolo di Chris Anderson chiamato "The Long Tail" trattò di questo argomento, affermando che l'industria fosse guidata proprio dai prodotti di nicchia. Misurare l'importanza della **long tail** è riconducibile all'analisi delle power laws.

### 18.5.1 Visualizing the Long Tail.

Nello studiare la long tail in relazione alle power laws ci accorgiamo da subito che la distribuzione degli elementi è nettamente diversa, e che il numero di elementi non popolari, messi insieme, ha una notevole importanza.

Per conciliare queste due visioni occorre modificare la definizione iniziale di popolarità, chiedendoci invece: come funzione di $k$, quale numero (invece che frazione) di elementi ha una popolarità di almeno (invece di esattamente) $k$?

Vediamo nella figura 72 una rappresentazione della nuova funzione; un punto $(k, j)$ in questa distribuzione significa che ci sono $j$ elementi che han venduto almeno $k$ copie.

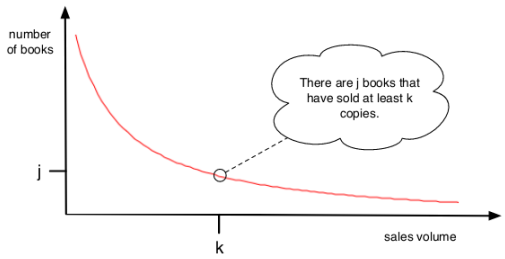



Figura 72: La distribuzione della popolarità nell'esempio dei libri: quanti oggetti hanno venduto almeno $k$ copie?.

Da questa visualizzazione possiamo semplicemente scambiare gli assi per poter osservare il volume di vendita degli oggetti meno venduti; in questo modo un punto $(k, j)$ ci dice che il $j$-esimo oggetto più popolare ha venduto $k$ copie. In questo modo, come si evince dalla figura 73, è più facile osservare l'effetto della long tail. In sostanza l'area sotto la curva da qualche punto $j$ in poi indica la quantità di vendite di oggetti di grado $j$ o maggiore.

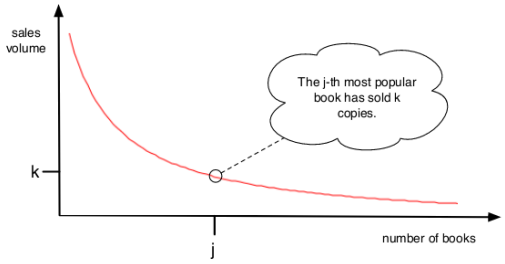



Figura 73: La distribuzione della popolarità nell'esempio dei libri: quante copie del $j$-esimo oggetto più popolare sono state vendute?

I grafici di questo tipo, con il rank invece della popolarità sull'asse $x$, sono spesso chiamati *Zipf plot*, dal nome di una linguista che produsse grafici di questo tipo nell'ambito di attività umane, formulando anche la legge di Zipf che afferma che nella lingua inglese la frequenza della $j$-esima parola più comune è proporzionale a $j$-1.

## 18.6 The Effect of Search Tools and Recommendation Systems

Tornando al discorso di Internet, gli strumenti di ricerca hanno aumentato o diminuito la dinamica richer-get-richer della popolarità? Abbiamo già visto

che un modello in cui le pagine copiano i link delle pagine precedenti alimenta questo aspetto, ma per quanto riguarda le persone che usano i motori di ricerca il discorso è diverso, poiché essi offrono già risultati in base alla popolarità, accentuandone ancora di più il livello. Ma dal momento che le possibili ricerche sono varie ed infinite e spesso sono legate agli interessi degli utenti, non c'è una singola lista di pagine migliori, e gli utenti sono portati a trovare pagine meno popolari ma più adatte a loro, contrastando l'effetti richer-get-richer.

Questa visione è parte integrante della discussione di Anderson sulla long tail, secondo cui per aumentare le vendite un'azienda deve offrire la possibilità ai clienti di esaminare tutto il catalogo in maniera ragionevole, come fanno i **recommendation systems** offerti da servizi quali Netflix o Amazon, che propongono i loro prodotti in base agli interessi dell'utente, che non per forza coincidono con le scelte più popolari.

## 18.7 Advanced Material: Analysis of Rich-Get-Richer Processes

Ci occupiamo ora di capire il comportamento di un modello che porta alla nascita delle power laws. Finora abbiamo descritto il modello come segue:

1. i nodi sono creati in ordine e chiamati $1, 2, \ldots, N$;
2. quando un nuovo nodo $j$ viene creato si connette ad una pagina precedente nel seguente modo:
   (a) con probabilità $p$, $j$ sarà connesso con una pagina $i$ casuale;
   (b) con probabilità 1 - $p$, $j$ sarà connesso con $l$ con una probabilità proporzionale al numero di in-links di $l$;
   (c) si ripete poi il processo per ogni nodo.

Abbiamo uno specifico processo casuale che itera per $N$ passi e possiamo stimare il numero di pagine con $k$ in-links alla fine del processo. Usiamo una approssimazione del modello che rende le cose un po' più semplici.

**A Deterministic Approximation of the Richer-Get-Richer Process.** Partendo dalle proprietà del modello, osserviamo che il numero di in-links di un nodo $j$ al tempo $t > j$ è una variabile casuale $X_j(t)$ che gode di queste due proprietà:

1. dato che il nodo $j$ è creato al tempo $j$ senza in-links, $X_j(j) = 0$;

2. il nodo $j$ ottiene in-links al tempo $t + 1$ se e solo se il nodo $t + 1$ punta ad esso. La probabilità che questo accada è $\frac{p}{t} + \frac{(1-p)X_j(t)}{t}$.

L'idea di base per analizzare un modello più semplice è di renderlo **deterministico**, cioè senza probabilità, ma che evolve nel tempo in un modo preciso. Il tempo scorre continuamente da 0 ad $N$ al posto di avere i passi discreti, e approssimiamo $X_j(t)$, il numero di in-links di $j$, come una funzione continua del tempo $x_j(t)$. Questa funzione ha due proprietà:

- inizialmente $x_j(0) = 0$;
- il fattore di crescita è dato dall'equazione differenziale $\frac{dx_j}{dt} = \frac{p}{t} + \frac{(1-p)x_j}{t}$.

**Solving the Deterministic Approximation.** Per risolvere l'equazione differenziale consideriamo $q$ = 1 - $p$, e procediamo in questo modo:

$$\frac{dx_j}{dt} = \frac{p}{t} + \frac{qx_j}{t}$$

dividiamo entrambi i membri per $p + qx_j$, ottenendo

$$\frac{1}{p+qx_j}\frac{dx_j}{dt} = \frac{1}{t}$$

integrando entrambe le parti

$$\int \frac{1}{p+qx_j}\frac{dx_j}{dt}dt = \int \frac{1}{t}dt$$

ottenendo

$$\ln(p + qx_j) = q \ln t + c$$

per una costante $c$. Usando l'esponente e scrivendo $A = e^c$ abbiamo:

$$p + qx_j = At^q$$

e quindi

$$x_j(t) = \frac{1}{q}(At^q - p)$$

Possiamo quindi determinare il valore di $A$ usando la condizione iniziale, che ci porta alla seguente equazione:

$$0 = x_j(j) = \frac{1}{q}(Aj^q - p)$$

e quindi $A = p/j^q$. Usando questo valore nell'equazione otteniamo:

$$x_j(t) = \frac{1}{q}(\frac{p}{j^q} \cdot t^q - p) = \frac{p}{q}[(\frac{t}{j})^q - 1]$$

**Identifying a Power Law in the Deterministic Approximation.** La ultima espressione ottenuta ci offre una visione della crescita di $x_j$ nel tempo. Usiamo questa informazione per capire per capire, dati un valore $k$ ed un tempo $t$, quale frazione di tutti i nodi ha almeno $k$ in-links al tempo $t$. Questo si traduce nel trovare tutte le $x_j$ tali che $x_j(t) \geq k$:

$$x_j(t) = \frac{p}{q}[(\frac{t}{j})^q - 1] \geq k$$

oppure riscrivendo in termini di $j$:

$$j \leq t[\frac{q}{p} \cdot k + 1]^{-1/q}$$

Di tutte le funzioni $x_1, x_2, \ldots, x_t$ al tempo t, la frazione di valori $j$ che soddisfa l'equazione è:

$$\frac{1}{t} \cdot t[\frac{q}{p} \cdot k + 1]^{-1/q} = [\frac{q}{p} \cdot k + 1]^{-1/q}$$

Cominciamo a vedere l'effetto della power law: dato che $p$ e $q$ sono costanti, l'espressione nelle parentesi quadre a destra è proporzionale a $k$, come la frazione di $x_j$ che è almeno $k$ è proporzionale a $k^{-1/q}$.

Infine passiamo dall'aver definito la frazione di nodi $F(k)$ con **almeno** $k$ in-links, all'approssimare la frazione $f(k)$ con **esattamente** $k$ in-links, semplicemente calcolando la derivata. Differenziando la precedente equazione otteniamo:

$$\frac{1}{q}\frac{q}{p}[\frac{q}{p} \cdot k + 1]^{-1-1/q}$$

In altre parole il modello deterministico prevede che la frazione di nodi con $k$ in-links è proporzionale a $k^{-(1+1/q)}$, una power law con esponente

$$1 + \frac{1}{q} = 1 + \frac{1}{1-p}$$

Quando $p$ tende ad 1, la formazione dei collegamenti è basata principalmente sul caso, e la dinamica richer-get-richer non entra in gioco. Al contrario, quando è tende a 0, la crescita del modello è strettamente legata a questa dinamica, e l'esponente della power law decresce a 2, permettendo diversi nodi con un grande numero di in-links.

# 6 Games

La **game theory** si occupa di analizzare le situazioni in cui il risultato della decisione di una persona dipende non solo dalla scelta tra diverse opzioni, ma anche dalle scelte degli altri individui nello stesso contesto, che può variare dai semplici giochi, alle aste, ad una strada in una trasportation network, ecc.

## 6.1 What Is a Game?

Per chiarire i primi aspetti della game theory partiamo con un esempio

### 6.1.1 A first Example

Pensiamo ad uno studente del college e che deve preparare due parti di un lavoro, un esame ed una presentazione, per il giorno seguente. Egli può solamente studiare per l'esame oppure preparare la presentazione, ma non può fare entrambe le cose. ed inoltre

Si assume di avere una precisa idea del risultato che si può ottenere: studiando per l'esame ci si aspetta un voto di 92, mentre non studiando un voto pari ad 80. La presentazione invece serve farla con un compagno, che però non si può contattare. Se entrambi preparano la presentazione si ottiene un voto di 100, se la prepara solamente uno un voto di 92, se nessuno la prepara un voto di 84.

Lo scopo è ovviamente di massimizzare il voto ricevuto, tenendo conto delle possibili scelte del compagno. I possibili scenari sono i seguenti:

- se entrambi preparano la presentazione, entrambi avranno 100 alla presentazione ed 80 all'esame, per una media di 90;
- se entrambi studiano per l'esame, entrambi prenderanno 92 all'esame e 84 alla presentazione, per una media di 88;
- se uno studia per l'esame e l'altro prepara la presentazione:
  - chi prepara la presentazione prende 92 per la presentazione e 80 per l'esame, per una media di 86;
  - chi studia per l'esame prende 92 per l'esame ma anche 92 per la presentazione, per una media di 92.

| | | Your Partner: *Presentation* | Your Partner: *Exam* |
|---|---|---|---|
| You | *Presentation* | $90, 90$ | $86, 92$ |
| | *Exam* | $92, 86$ | $88, 88$ |

Figura 74: La tabella delle diverse opzioni nell'esempio esame-presentazione.

Possiamo riassumere i diversi casi nella tabella della figura 74. Passiamo ora ad alcune definizioni base della teoria dei giochi.

### 6.1.2 Basic Ingredients of a Game

La situazione appena descritta è una classico esempio di **gioco**, ovvero una situazione con i seguenti aspetti:

1. c'è un insieme di partecipanti, chiamati **players**;
2. ogni giocatore ha una scelta di possibili opzioni per come comportarsi. Ci riferiremo ad esse come possibili **strategies**;
3. per ogni scelta di strategia, ogni giocatore riceve un **payoff** che può dipendere anche dalle strategie altrui. Generalmente il guadagno è un numero e ogni giocatore preferisce un guadagno maggiore. La figura 74 viene chiamata **payoff matrix**.

Il nostro interesse è ragionare su come si comporterebbero i giocatori in un dato gioco.

## 6.2 Reasoning about Behavior in a Game

Avendo descritto le basi di un gioco, passiamo ad occuparci del comportamento dei giocatori.

### 6.2.1 Underlying Assumptions

Cominciamo con alcune assunzioni. La prima è che tutto ciò che importa ad un giocatore è riassunto nella matrice payoff. Se ad esempio ad un giocatore importa di massimizzare anche il payoff degli altri oltre al proprio, questo deve essere rappresentato direttamente nella matrice. Assumiamo inoltre che ogni giocatore conosca ogni particolare della struttura del gioco, dai partecipanti alle possibile scelte proprie ed altrui. Infine supponiamo ancora che ogni giocatore scelga una strategia per massimizzare il proprio payoff, secondo ciò

che pensa facciano gli altri partecipanti. Questo modello di comportamento è noto come **rationality**, e comprende il fatto che ogni giocatore faccia la scelta migliore. Altri tipi di giochi prevedono invece un approccio diverso, in cui il giocatore può sbagliare ed imparare, migliorando la mossa successiva, ma sono scenari più complessi da studiare.

### 6.2.2 Reasoning about Behaviour in the Exam-or-Presentation Game

Tornando all'esempio dell'esame o presentazione, ci concentriamo sulle possibili scelte da effettuare conoscendo la scelta dell'altro giocatore:

- sapendo che il partner studierà per l'esame, la mossa migliore sarebbe studiare per l'esame per un payoff di 88;
- sapendo invece che il partner preparerà la presentazione, la mossa migliore sarebbe nuovamente studiare, per un payoff di 92.

Questo ragionamento ci porta a pensare che, a prescindere da cosa farà il partner, la scelta migliore è quella di studiare per l'esame; quando come in questo caso una strategia è strettamente migliore delle altre, si parla di **strictly dominant strategy**.

Nonostante ciò però notiamo che se ci fosse la possibilità di mettersi d'accordo si potrebbero raggiungere risultati migliori; ma se un giocatore decidesse di preparare la presentazione, sperando di ricevere entrambi un voto di 90, l'altro giocatore avrebbe l'incentivo di studiare per l'esame, per massimizzare il proprio payoff. Questo perché il payoff è basato, da assunzione, sul proprio guadagno personale, e non tiene conto del risultato ottenuto dagli altri.

### 6.2.3 A Related Story: The Prisoner's Dilemma

Un altro famoso esempio di gioco, simile al precedente, è quello noto con il nome di **prisoner's dilemma**. Esso funziona nel seguente modo: supponiamo che due sospettati siano stati arrestati dalla polizia e vengano interrogati in stanze separate. La polizia pensa che entrambi siano complici di una rapina, ma non ci sono prove evidenti per condannarli. In ogni caso entrambi hanno fatto resistenza all'arresto, quindi la pena minima sarebbe di 1 anno. Ad ogni sospettato viene detto che se egli confesserà ed il partner non farà altrettanto, allora verrà liberato ed il partner arrestato e condannato per 10 anni. Se entrambi confesseranno invece la pena sarà di 4 anni.

Analizziamo il payoff nella matrice della figura seguente:

| | | Suspect 2 | |
|---|---|---|---|
| | | $NC$ | $C$ |
| Suspect 1 | $NC$ | $-1, -1$ | $-10, 0$ |
| | $C$ | $0, -10$ | $-4, -4$ |

Figure 6.2: Prisoner's Dilemma

Figura 75: La matrice payoff del prisoner's dilemma.

Come nel caso precedente analizziamo le diverse opzioni di ogni sospettato:

- se l'altro sospettato confessasse, allora l'opzione migliore sarebbe confessare, per un payoff di -4;
- se l'altro sospettato non confessasse, l'opzione migliore sarebbe di confessare per un payoff di 0.

Quella di confessare sarebbe quindi la strategia strettamente dominante, e come risultato ci si aspetta quindi che entrambi confessino, per un payoff comune di -4.

Come nel caso precedente il risultato migliore per entrambi, ovvero il caso in cui nessuno confessi, non viene considerato dalla strategia più razionale.

### 6.2.4 Interpretations of the Prisoner's Dilemma

Per la sua semplicità ma allo stesso tempo capacità di esprimere e rappresentare diversi aspetti della game theory, il prisoner's dilemma è stato oggetto di studio per diversi anni.

Un esempio simile è quello dell'uso di doping nello sport, dove i partecipanti sono gli atleti e le possibili strategie sono l'usare o meno le droghe. Facendone uso, si ottiene un vantaggio nella competizione, ma con effetti negativi nel tempo. La figura seguente indica la matrice payoff di questo esempio:

| | | Athlete 2 | |
|---|---|---|---|
| | | *Don't Use Drugs* | *Use Drugs* |
| Athlete 1 | *Don't Use Drugs* | $3, 3$ | $1, 4$ |
| | *Use Drugs* | $4, 1$ | $2, 2$ |

Figura 76: La matrice payoff dell'esempio degli atleti.

Anche in questo caso c'è una strategia strettamente dominante, quella di usare doping, ma esiste una opzione migliore che i giocatori non considerano, cioè quella che nessuno ne faccia uso.

Più in generale le situazioni di questo tipo sono chiamate **arms races**, in cui i partecipanti usano strategie pericolose per poi avere entrambi un ugual vantaggio, ma con un risultato peggiore di quello ottenibile quando nessuno fa uso di queste strategie.

Tornando al prisoner's dilemma, possiamo aggiungere che queste situazioni sono spessi comuni in casi di payoff bilanciati, ma il risultato può cambiare drasticamente in caso di valori non equi. Ad esempio la figura 77 mostra una variante del problema dell'esame, in cui la presentazione è uguale, ma l'esame è più semplice, perciò studiando si ottiene un voto pari a 100, mentre se non si studia si ottiene 96. In questo caso preparare la presentazione diventa una strategia strettamente dominante, senza i lati negativi del caso precedente.

| | | Your Partner | |
|---|---|---|---|
| | | *Presentation* | *Exam* |
| You | *Presentation* | $98, 98$ | $94, 96$ |
| | *Exam* | $96, 94$ | $92, 92$ |

Figura 77: Un'altra versione dellla matrice payoff dell'esempio esame-presentazione.

## 6.3 Best Responses and Dominant Strategies

In questi esempi abbiamo introdotto due concetti chiave. Il primo è l'idea di **best response**, ossia la miglior scelta per un giocatore, date le possibili mosse degli altri. Se $S$ è la strategia scelta dal Giocatore 1, e $T$ è la strategia scelta dal Giocatore 2, allora c'è una cella nella matrice payoff che corrisponde alle due strategia ($S$, $T$). Indichiamo con $P_1(S, T)$ il payoff del Giocatore 1 per questa coppia di strategie, e con $P_2(S, T)$ il payoff corrispondente del Giocatore 2. Diciamo che $S$ è la best response se per qualsiasi altra strategia $S'$ vale:

$$P_1(S, T) \geq P_1(S', T)$$

Lo stesso discordo ovviamente vale per il Giocatore 2 e la strategia da lui adottata.

Si noti che questa definizione ammette più strategie possibili in risposta alla strategia $T$, per questo possiamo definire la **strict best response** come:

$$P_1(S, T) > P_1(S', T)$$

Il secondo concetto è legato alla strategia strettamente dominante, che si può formulare in questo modo:

- diciamo che una **dominant strategy** per il Giocatore 1 è una strategia che è una best response per ogni strategia del Giocatore 2;
- diciamo che una **strictly dominant strategy** per il Giocatore 1 è una strategia che è una strict best response per ogni strategia del Giocatore 2.

Negli esempi precedenti la strategia era sempre molto chiara perché tutti presentava una strategia strettamente dominante, ma ora passiamo a casi più complessi in cui la situazione necessità di analisi più complesse.

### 6.3.1 A Game in Which Only One Player Has a Strictly Dominant Strategy

Inizialmente consideriamo un contesto in cui solo un giocatore ha una strategia strettamente dominante. Prendiamo come esempio due aziende che devono vendere un nuovo prodotto, che può essere diviso in due segmenti di mercato: le persone che comprerebbero solo una versione economica del prodotto e le persone che ne comprerebbero solo la versione costosa. Ogni azienda vuole ovviamente massimizzare i propri guadagni, e per questo deve decidere se produrre la versione economica o quella costosa. Per determinare il payoff facciamo le seguenti considerazioni:

- le persone che preferiscono la versione economica sono il 60%, quelli che preferiscono quella costosa il 40%;
- l'Azienda 1 è molto più popolare, perciò quando entrambe le aziende competono nello stesso settore essa prende l'80% delle vendite, mentre l'Azienda 2 il 20%.

Con questi presupposti possiamo determinare il payoff:

- se le due aziende hanno segmenti di mercato differenti, ottengono i pieni ricavi per il loro settore, ovvero il 60% per un payoff di .60 per quella che vende il prodotto economico, ed il 40% con un payoff di .40 per quella che vende quello costoso;
- se entrambe le aziende vendono il prodotto economico, l'Azienda 1 ottiene l'80% del segmento per un payoff di .48, mentre l'Azienda 2 il 20% per un payoff di .12;

- se entrambe le aziende vendono il prodotto costoso, l'Azienda 1 ottiene un payoff di .32 e l'Azienda 2 un payoff di .08.

I valori di payoff sono riassunti nella figura 78.

| | | Firm 2 | |
|---|---|---|---|
| | | *Low-Priced* | *Upscale* |
| Firm 1 | *Low-Priced* | .48, .12 | .60, .40 |
| | *Upscale* | .40, .60 | .32, .08 |

Figure 6.5: Marketing Strategy

Figura 78: La matrice payoff per l'esempio delle aziende.

In questo gioco l'Azienda 1 ha una strategia strettamente dominante perché il prodotto economico è una strict best response per ogni strategia dell'altra azienda. L'Azienda 2 invece non ha una strategia dominante, perché il prodotto economico è la best response nel caso in cui l'Azienda 1 venda il prodotto costoso, e viceversa. Ci aspettiamo quindi che l'Azienda 1 attui la strategia migliore, e che l'Azienda 2 agisca di conseguenza, ma dobbiamo ricordare che entrambe le mosse vengono eseguite contemporaneamente.

Nel gioco si suppone che tutti i partecipanti conoscano la matrice payoff e che ognuno cerchi di massimizzare il proprio payoff, e che sia nota a tutti la struttura del gioco. Questo equivale a dire che i partecipanti hanno una **common knowledge** del gioco.

## 6.4 Nash Equilibrium

Quando nessun giocatore ha una strategia strettamente dominante, serve prevedere in qualche altra maniera l'esito del gioco.

### 6.4.1 An Example: A Three-Client Game

Modifichiamo l'esempio delle due aziende per analizzare questo nuovo contesto, aggiungendo la possibilità per ogni azienda di fare affari con uno di tre clienti A, B, e C. I possibili risultati delle scelte sono i seguenti:

- se le due aziende si approcciano allo stesso cliente, si spartiscono equamente gli affari;
- l'Azienda 1 è troppo piccola per fare affari da sola, perciò se decide di approcciare un cliente diverso dall'Azienda 2 ottiene un payoff di 0;

- se l'Azienda 2 approccia il cliente B o C da sola, ottiene l'intero mercato. Il cliente A invece è molto più grande e accetta di fare affari solamente con entrambe le aziende contemporaneamente;
- poiché A è un cliente molto grande, fare affari con esso ha valore 8, mentre fare affari con B o C ha valore pari a 2.

Possiamo vedere la matrice payoff di questo esempio nella seguente figura.

| | | Firm 2 | | |
|---|---|---|---|---|
| | | $A$ | $B$ | $C$ |
| Firm 1 | $A$ | $4,4$ | $0,2$ | $0,2$ |
| | $B$ | $0,0$ | $1,1$ | $0,2$ |
| | $C$ | $0,0$ | $0,2$ | $1,1$ |

Figura 79: La matrice payoff per l'esempio delle aziende con tre clienti.

Studiando la matrice vediamo che nessuna delle due aziende ha una strategia dominante e che ogni azienda ha una strict best response ad una strategia dell'altro.

### 6.4.2 Defining Nash Equilibrium

Nel 1950 John Nash propose una semplice ma potente soluzione per risolvere questo quesito, partendo dall'idea che anche in caso non ci siano strategie migliori, ogni giocatore tende ad usare la strategie che risponde meglio a quella avversaria.

Supponendo che il Giocatore 1 scelga la strategia $S$ ed il Giocatore 2 la strategia $T$, diciamo che questa coppia di strategie è un **Nash equilibium** se $S$ è la best response a $T$ e viceversa. L'idea nasce dal concetto stesso di equilibrio, e dal fatto che se ogni giocatore sceglie come strategia una best response verso l'altro, nessuno ha l'incentivo di cambiare.

Due strategie che invece non sono una best response reciproca non costituiscono un equilibrio, perché entrambi i giocatori sanno che queste strategie non verranno usate, dal momento che almeno uno di essi avrà un incentivo a cambiare la propria. Il Nash equilibrium può quindi essere visto come un equilibrio nelle convinzioni, perché se ogni giocatore crede che l'altro abbia l'incentivo di usare la strategia parte del Nash equilibrium, anche egli sarà incentivato a fare altrettanto.

Consideriamo il gioco dei tre clienti dalla prospettiva del Nash equilibrium. Se entrambe le aziende scelgono A, sappiamo che ognuna sceglie la best response alla strategia altrui, formando una situazione di equilibrio.

Controllando le altre opzioni, notiamo che la coppia (A, A) è l'unica che forma il Nash equilibrium. Possiamo quindi cercare l'equilibrio controllando tutte le possibili coppie di strategie e chiedendoci per ognuna se la le strategie sono reciprocamente best responses; oppure possiamo cercarlo calcolando la best response di ogni giocatore per ogni strategia dell'altro e considerare le strategie che sono best responses reciproche.

## 6.5 Multiple Equilibria: Coordination Games

Alcuni giochi ammettono però più di una situazione di Nash equilibrium ed in questi casi è difficile prevedere il comportamento di un giocatore.

### 6.5.1 A Coordination Game

Un esempio semplice ma importante è quello del **coordination game**: supponiamo che due colleghi stiano preparando una presentazione per un lavoro, senza poter comunicare. Si può scegliere di preparare la presentazione su Powerpoint o su Keynote, ma scegliere lo stesso software rende la parte di unione delle slides più semplice. La matrice payoff di questo esempio è rappresentata nella figura seguente

| | | Your Partner | |
|---|---|---|---|
| | | *PowerPoint* | *Keynote* |
| You | *PowerPoint* | $1, 1$ | $0, 0$ |
| | *Keynote* | $0, 0$ | $1, 1$ |

Figura 80: La matrice payoff per l'esempio della presentazione.

Questo tipo di scenario è chiamato coordination game perché l'obiettivo è appunto di coordinarsi sulla stessa strategia. La difficoltà in questo caso è data dal fatto che esistono due Nash equilibria. Thomas Schelling ha introdotto l'idea del **focal point** per risolvere questo problema, affermando che in diversi giochi c'è una ragione naturale, come ad esempio le convenzioni sociali, che porta i giocatori a concentrarsi su un preciso equilibrio.

### 6.5.2 Variants on the Basic Coordination Game

Estendiamo il gioco della presentazione ipotizzando che uno dei due giocatori preferisca Powerpoint e l'altro Keynote. C'è ancora la volontà di coordinarsi, ma ora le due alternative non sono uguali. La figura 81 mostra la matrice payoff per questo caso di **unbalanced coordination game**.

| | | Your Partner | |
|---|---|---|---|
| | | *PowerPoint* | *Keynote* |
| You | *PowerPoint* | $1,1$ | $0,0$ |
| | *Keynote* | $0,0$ | $2,2$ |

Figura 81: La matrice payoff per l'esempio della presentazione non bilanciato.

I Nash equilibria sono gli stessi, ma danno payoff divesi ai giocatori. Schelling in questo caso suggerisce che si può usare una caratteristica intrinseca del gioco, invece di una convenzione sociale, per fare una predizione sull'equilibrio che verrà scelto, ovvero quello che porta un maggiore payoff ad entrambi.

La situazione è più complessa nella matrice della figura 82, in cui l'equilibrio è portato dalle stesse scelte, ma il payoff è distribuito diversamente tra i giocatori, a seconda della scelta. Questo scenario è tipicamente chiamato **battle of the sexes**.

| | | Your Partner | |
|---|---|---|---|
| | | *PowerPoint* | *Keynote* |
| You | *PowerPoint* | $1,2$ | $0,0$ |
| | *Keynote* | $0,0$ | $2,1$ |

Figura 82: La matrice payoff per l'esempio della battle of sexes.

Un'ultima variazione del coordination game è lo stag hunt game, in cui i due giocatori possono cacciare un cervo solamente se collaborano, altrimenti possono cacciare una lepre per conto proprio. La situazione è simile al coordination game non bilanciato, con la differenza che senza coordinazione il giocatore che ha provato ad avere il payoff massimo non ottiene nulla. In questo senso è molto simile al prisoner's dilemma, con la differenza però che quest'ultimo presenta delle strategie strettamente dominanti. Possiamo anche modificare il gioco esame-presentazione per renderlo simile allo stag hunt; vediamo di seguito le matrici payoff del gioco stag hunt e dell'esame-presentazione modificato.

| | | Hunter 2 | |
|---|---|---|---|
| | | *Hunt Stag* | *Hunt Hare* |
| Hunter 1 | *Hunt Stag* | $4,4$ | $0,3$ |
| | *Hunt Hare* | $3,0$ | $3,3$ |

Figura 83: La matrice payoff per l'esempio del gioco stag hunt.

| | | Your Partner: *Presentation* | Your Partner: *Exam* |
|---|---|---|---|
| You | *Presentation* | 90, 90 | 82, 88 |
| | *Exam* | 88, 82 | 88, 88 |

Figura 84: La matrice payoff per l'esempio del gioco esame-presentazione modificato.

## 6.6 Multiple Equilibria: The Hawk-Dove Game

Si possono trovare più Nash equilibria anche nei giochi di tipo **anticoordination**, come ad esempio il gioco **hawk-dove**: supponiamo che due animali debbano decidere come dividersi il cibo; ogni animale può essere aggressivo (strategia hawk) o passivo (dove strategy). Se entrambi si comportano passivamente, il cibo viene diviso equamente, se scelgono strategie diverse l'aggressivo ne ottiene di più, mentre se entrambi si comportano aggressivamente il cibo viene distrutto. La figura seguente riassume queste regole nella matrice payoff.

| | | Animal 2: *D* | Animal 2: *H* |
|---|---|---|---|
| Animal 1 | *D* | 3, 3 | 1, 5 |
| | *H* | 5, 1 | 0, 0 |

Figura 85: La matrice payoff per l'esempio del gioco hawk-dove.

Questo gioco ha come Nash equilibria le scelte (D, H) e (H, D). Come nel caso dei coordination games il concetto di Nash equilibrium restringe le scelte possibili, ma ne offre comunque più di una, perciò prevedere l'esito del gioco è difficile.

La figura 86 mostra invece una versione del gioco esame-presentazione, modificata per rispecchiare le caratteristiche del gioco hawk-dove

| | | Your Partner: *Presentation* | Your Partner: *Exam* |
|---|---|---|---|
| You | *Presentation* | 90, 90 | 86, 92 |
| | *Exam* | 92, 86 | 76, 76 |

Figura 86: La matrice payoff per l'esempio del gioco esame-presentazione nello stile hawk-dove.

## 6.7 Mixed Strategies

Fino ad ora abbiamo visto giochi che ammettevano più Nash equilibria, ma esistono anche giochi che non ne ammettono nessuno: per questo tipo di giochi ammettiamo delle strategie casuali. Il tipo migliore di giochi per descrivere questo aspetto sono gli **attack-defense**, dove l'attaccante sceglie una strategia A o B, ed il difensore scegliere di difendersi contro A oppure B.

### 6.7.1 Matching Pennies

Un semplice gioco di questo tipo è il **matching pennies**, in due giocatori scelgono se mostrare testa (H) o croce (T) dalla loro moneta. Il Giocatore 1 vince se il lato mostrato dai due giocatori è diverso. Questo produce la matrice payoff mostrata di seguito.

| | | Player 2 $H$ | Player 2 $T$ |
|---|---|---|---|
| Player 1 | $H$ | $-1, +1$ | $+1, -1$ |
| | $T$ | $+1, -1$ | $-1, +1$ |

Figura 87: La matrice payoff per l'esempio del gioco matching pennies.

La particolarità del gioco matching pennies è che la somma dei payoff per ogni coppia di scelte è pari a zero; questo tipo di giochi è chiamato **zero-sum**, e molti attack-defense hanno questa struttura.

Notiamo che non esiste una coppia di strategie che sono una best response reciproca, ed il giocatore è spinto a muoversi verso un payoff maggiore. Non abbiamo quindi Nash equilibrium in questo scenario. Se pensiamo a come questi giochi sono affrontati nella vita reale, pensiamo che la mossa migliore sia rendere imprevedibile la propria mossa. Per questo è una buona idea introdurre un fattore di casualità nel modello del gioco.

### 6.7.2 Mixed Strategies

Il modo migliore per introdurre l'elemento stocastico è che ogni giocatore scelga una probabilità con cui attuerà la strategia H o T. Con probabilità $p$ il Giocatore 1 sceglierà la mossa H, e con probabilità 1 - $p$ sceglierà la mossa T. Lo stesso vale per il Giocatore 2 con la probabilità $q$.

Dal momento che le regole ed il contesto del gioco sono cambiati, poiché non si tratta più di due strategie per ogni giocatore, bensì di un insieme di strategie corrispondente all'intervallo di numeri tra 0 ed 1, si parla di **mixed**

**strategies**. Le strategie originali T ed H corrispondono ai valori 0 ed 1, e vengono chiamate **pure strategies**.

### 6.7.3 Payoffs from Mixed Strategies

Con questi insiemi di strategie dobbiamo anche determinare i payoffs. Serve quindi un principio per determinare quando un risultato casuale è migliore di un altro, e per ordinare i payoffs casuali numericamente assegniamo un numero ad ogni distribuzione, in modo da rappresentare quanto questa distribuzione sia attraente per il giocatore. Quando abbiamo assegnato questi numeri, possiamo ordinare le distribuzioni, usando come numeri loro associati gli **expected values** dei payoffs.

Ad esempio se il Giocatore 1 sceglie la strategia pura H e il Giocatore 2 sceglia la probabilità $q$, abbiamo che il payoff atteso dal Giocatore 1 è:

$$(-1)q + (1)(1-q) = 1 - 2q$$

Assumiamo che ogni giocatore cerchi di massimizzare il payoff atteso dalla scelta di strategie miste.

### 6.7.4 Equilibrium with Mixed Strategies

La definizione di equilibrio in questo sistema arricchito è la stessa vista in precedenza, e sappiamo che nel gioco matching pennies originale non si può raggiungere il nash equilibrium. Analizzando questo caso invece, ipotizziamo che per il Giocatore 1 la strategia pura H sia parte di Nash equilibrium; allora la risposta migliore per il Giocatore 2 sarebbe sempre H, ma H per il Giocatore 1 non sarebbe la risposta migliore alla H del Giocatore 2, perciò non può esserci equilibrio. Serve quindi una strategia compresa tra 0 ed 1, estremi esclusi.

In precedenza abbiamo detto che il payoff del Giocatore 1 con la strategia H in risposta alla strategia $q$ del Giocatore 2 è 1 - $2q$, mentre il payoff per la strategia T è $2q$ - 1. Se questi due valori fossero diversi, una sola strategia pura sarebbe la best risponse alla strategia $q$ del Giocatore 2. Ma in questo caso uno dei due valori è per forza maggiore dell'altro, perciò non avrebbe senso che il Giocatore 1 puntasse sulla strategia più debole. Abbiamo anche detto che le strategie pure non sono parte del Nash equilibrium, e dato che esse sono la best response quando $1 - 2q \neq 2q - 1$, allora le probabilità che rendono questi valori diversi non sono ammesse.

Otteniamo quindi che in ogni scenario del gioco matching pennies per avere Nash equilibrium serve che $q$ valga $1/2$, ed in conseguenza anche $p$. La situazione è simmetrica per il Giocatore 2.

### 6.7.5 Interpreting the Mixed-Strategy Equilibrium for Matching Pennies

Avendo derivato il Nash equilibrium per questo gioco, possiamo ora generalizzarlo per ogni contesto. Il porre $q$ pari ad $1/2$ nell'esempio precedente, rende di fatto la mossa **indifferente** per il giocatore, che risulta quindi imprevedibile.

La nozione di indifferenza è un principio generale nel calcolo del Nash equilibrium per le strategie miste, perché non permette di prevedere e quindi contrastare una determinata strategia.

## 6.8 Mixed Strategies: Examples and Empirical Analysis

Usiamo altri due esempi per capire meglio l'equilibrio nelle strategie miste.

### 6.8.1 The Run-Pass Game

Questo è l'esempio di un incontro di football americano. La squadra che attacca può scegliere di correre o passare, mentre la squadra che difende può scegliere di difendere contro la corsa o contro il passaggio. Il payoff funziona in questo modo:

- se la difesa intercetta la mossa giusta, l'attacco ottiene 0 metri;
- se l'attacco corre e la difesa difende contro il passaggio, l'attacco ottiene 5 metri;
- se l'attacco passa e la difesa difende contro la corsa, l'attacco ottiene 10 metri.

Abbiamo quindi la seguente matrice:

| | | Defense | |
|---|---|---|---|
| | | *Defend Pass* | *Defend Run* |
| Offense | *Pass* | $0, 0$ | $10, -10$ |
| | *Run* | $5, -5$ | $0, 0$ |

Figura 88: La matrice payoff per l'esempio del gioco run-pass.

Vediamo chiaramente che non c'è Nash equilibrium usando strategie pure, perciò i due giocatore devo introdurre scelte casuali: sia $p$ la probabilità che l'attacco passi, e $q$ la probabilità che la difesa difenda contro il passaggio.

In questo caso sappiamo che ci sarà un equilibrio, ma non per quali valori delle due probabilità. Il payoff invece varrà:

- se la difesa sceglie la probabilità $q$ per la difesa dal passaggio, il payoff atteso per l'attacco che decide di passare sarà: $(0)(q) + (10)(1-q) = 10 - 10q$. Il payoff per l'attacco che decide di correre sarà invece: $(5)q + (0)(1-q) = 5q$. Per rendere le due strategie indifferenti per l'attacco dobbiamo porre $10 - 10q = 5q$
  $rightarrow q = 2/3$;

- se l'attacco sceglie la probabilità $p$ per passare, seguendo lo stesso ragionamento la difesa avrà un payoff pari a $5p - 5$ per la difesa dal passaggio, ed un payoff pari a $-10p$ per la difesa dalla corsa. Troviamo quindi che $5p - 5 = -10p \rightarrow p = 1/3$.

Abbiamo calcolato quindi i valori delle due probabilità che portano ad un Nash equilibrium nel sistema, con un payoff di 10/3 per l'attacco e -10/3 per la difesa.

### 6.8.2 Strategic Interpretation of the Run-Pass Game

Notiamo che nonostante il passare sia la mossa più potente per l'attacco, essa è usata meno della metà delle volte, poiché la difesa tenderebbe a difendere più volte contro quella mossa. Se infatti usiamo valori più alti di $p$, come $1/2$, sappiamo che la difesa difenderebbe sempre contro il passaggio, ed il payoff per l'attacco sarebbe $5/2$, minore di quanto atteso in precedenza.

In equilibrio, la potenza della mossa è espressa dal fatto che la difesa difende dal passaggio $2/3$ delle volte, nonostante esso venga usato solo $1/3$ delle volte. Questo indica il livello di **threat** della mossa, e le naturali risposte della difesa.

### 6.8.3 The Penalty-Kick Game

Vediamo ora un altro esempio, sempre dal mondo dello sport: quello dei calci di rigore nel calcio. L'attaccante può tirare a destra o a sinistra, ed il portiere può decidere di buttarsi da una delle due parti. Nel 2002 Ignacio Palacios-Huerta studio a lungo questo scenario osservando circa 1400 calci di rigore, e considerando fattori come il fatto che la maggior parte dei calciatori è destrorsa e che si può segnare anche se il portiere su tuffa nella giusta direzione, stilò la seguente matrice payoff:

| | | Goalie $L$ | Goalie $R$ |
|---|---|---|---|
| Kicker | $L$ | $0.58, -0.58$ | $0.95, -0.95$ |
| | $R$ | $0.93, -0.93$ | $0.70, -0.70$ |

Figura 89: La matrice payoff per l'esempio del gioco penalty kick.

Usando lo stesso principio del precedente esempio, scegliamo $q$ come la probabilità che il portiere si tuffi a sinistra, ed otteniamo che $q = .42$. Con lo stesso calcolo per rendere il portiere indifferente alla scelta, otteniamo $p = .39$, dei valori che sono molto vicini a quelli ottenuti dagli studi.

### 6.8.4 Finding All Nash Equilibria

Per concludere il discorso, è importante notare che un gioco può avere un equilibrio puro oppure misto. Perciò per analizzare il Nash equilibrium conviene prima provare le strategie pure, e dopo inoltrarsi in quelle miste, trovando le probabilità $p$ e $q$. Abbiamo visto questo processo in diversi giochi di tipo attacco-difesa, ma possiamo trovare un mixed equilibrium anche nei giochi di coordinazione come quello esame-presentazione non bilanciato, la cui matrice payoff è mostrata nella figura 90.

| | | Your Partner *PowerPoint* | Your Partner *Keynote* |
|---|---|---|---|
| You | *PowerPoint* | $1, 1$ | $0, 0$ |
| | *Keynote* | $0, 0$ | $2, 2$ |

Figura 90: La matrice payoff per l'esempio del gioco esame-presentazione non bilanciato.

Supponiamo di avere una probabilità $p$ per il Giocatore 1 e la probabilità $q$ per il Giocatore 2 di scegliere Powerpoint, il Giocatore 1 è quindi indifferente nella scelta se $(1)(q)+(0)(1-q) = (0)(q)+(2)(1-q) \rightarrow q = 2/3$. La situazione è simmetrica perciò anche $p = 2/3$, ottenendo così un altro equilibrio oltre a quello delle strategie pure.

## 6.9 Pareto Optimality and Social Optimality

In una situazione di Nash equilibrium ogni strategia di ogni giocatore è la migliore risposta alla strategia dell'altro, ovvero i giocatori sono in una situazione ottimale individuale. Ma questo non significa che i giocatori arrivino ad un risultato ottimo, come nel caso dell'esame-presentazione o del

prisoner's dilemma visti in precedenza. Vediamo quindi due definizioni per rappresentare questo problema.

### 6.9.1 Pareto Optimality

La prima definizione è quella di **Pareto-optimality**, dal nome di Vilfredo Pareto, definita come:

> Una scelta di strategie, una per ogni giocatore, è una Pareto-optimality se non c'è un'altra scelta di strategie in cui tutti i giocatori ricevono un payoff almeno così alto, ed almeno un giocatore riceve un payoff strettamente più alto.

Quindi se una scelta non è Pareto-optimal c'è una scelta che porta almeno un giocatore ad un payoff più alto, senza danneggiare gli altri giocatori, ovvero è intuitivamente una scelta migliore. I giocatori potrebbero accordarsi per usare la strategia migliore per tutti, che sarebbe anch'essa ottima, altrimenti almeno un giocatore non vorrebbe utilizzarla. Il problema però è che oltre a realizzare che ci sia una strategia migliore per tutti, bisogna mettersi di comune accordo per sceglierla, e ciò in diversi scenari non è fattibile.

### 6.9.2 Social Optimality

Una condizione più forte, ma al contempo più semplice da definire, è la **social optimality**:

> Una scelta di strategie, una per ogni giocatore, è un **social welfare maximizer** (o socially optimal) se massimizza la somma dei payoffs dei giocatori.

I risultati che sono social optimal sono anche Pareto-optimal, perchè se non lo fosse ci sarebbe un'altra scelta in cui i payoffs sarebbero uguali, con almeno uno più grande di prima, e questo produrrebbe una somma maggiore. Un risultato Pareto-optimal non è però per forza social optimal .

# Elenco delle figure