# 3 Network Science

Sommario

[3 Network Science 1](#_Toc11079838)

[Teora dei grafi (o reti in network science) 2](#_Toc11079839)

[Grafi indiretti 2](#_Toc11079840)

[Grafo diretto 3](#_Toc11079841)

[Grafi semplici e grafi multipli 4](#_Toc11079842)

[Grafi pesati 4](#_Toc11079843)

[Cammino 5](#_Toc11079844)

[Ciclo 5](#_Toc11079845)

[Cammino minimo 5](#_Toc11079846)

[Lunghezza 6](#_Toc11079847)

[Grafi connessi e fortemente connessi 6](#_Toc11079848)

[Componenti connesse e fortemente connesse 6](#_Toc11079849)

[Grafo bipartito 6](#_Toc11079850)

[Grado di un nodo 7](#_Toc11079851)

[Grado pesato di un nodo 8](#_Toc11079852)

[Network science in R - A tidy approach 9](#_Toc11079853)

[1) Misure centralità dei nodi 10](#_Toc11079854)

[2) Misura centralità per archi 12](#_Toc11079855)

[3) Misura di similarità fra coppie di nodi 13](#_Toc11079856)

[4) Clustering gerarchico 14](#_Toc11079857)

[Visualizzazione interattiva delle reti con Visnetwork 16](#_Toc11079858)

[Esercizio in classe sui delfini 16](#_Toc11079859)

registrazione 1315 (tutorial sui grafi)

Non tutti i tipi di dati hanno una forma tabellare, e quindi esulano dall’universo tidyverse. Vediamo 3 importanti esempi che vedremo nel seguito del corso.

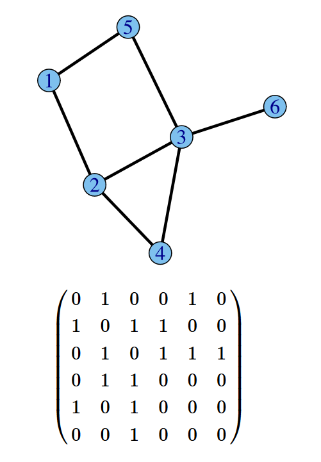
* **Reti**: (i grafi) possono essere rappresentate attraverso un DFrame per i nodi e DFrame per gli archi. Ma tipicamente è usata la *matrice di adiacenza*.
* **Dati gerarchici** : chiamati *dati semi-strutturati* cioè hanno una si una struttura non così rigida da essere incasselata in una tabella. Ad esempio, documento xml, dati html. I tag danno una semistruttura, ma non una tabella (non possiamo mettere una pagina html in una tabella).
* **Testo**: un libro non c’entra niente con una tabella, anche se vedremo un approccio che si basa sulla struttura del DFrame per fare l’analisi del testo.

Dovremo per cui analizzarli con un approccio a se stante rispetto il tidy.

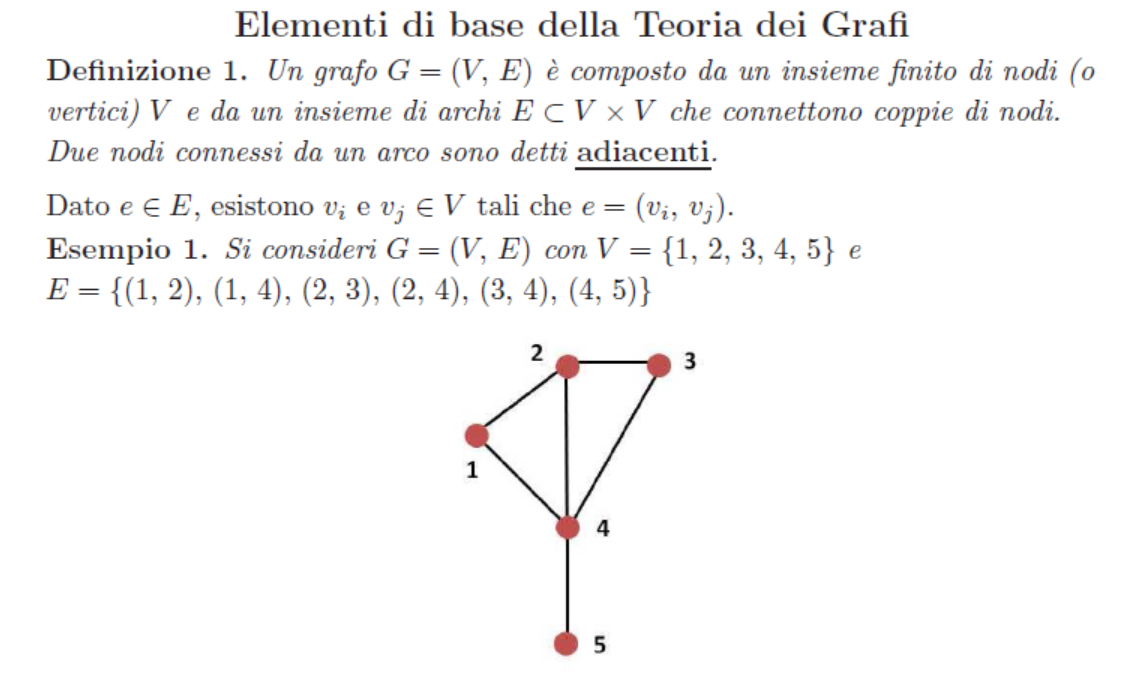
Ripasso sui grafi, tratto da <http://users.dimi.uniud.it/~massimo.franceschet/teaching/datascience/network/graphtheory.html>

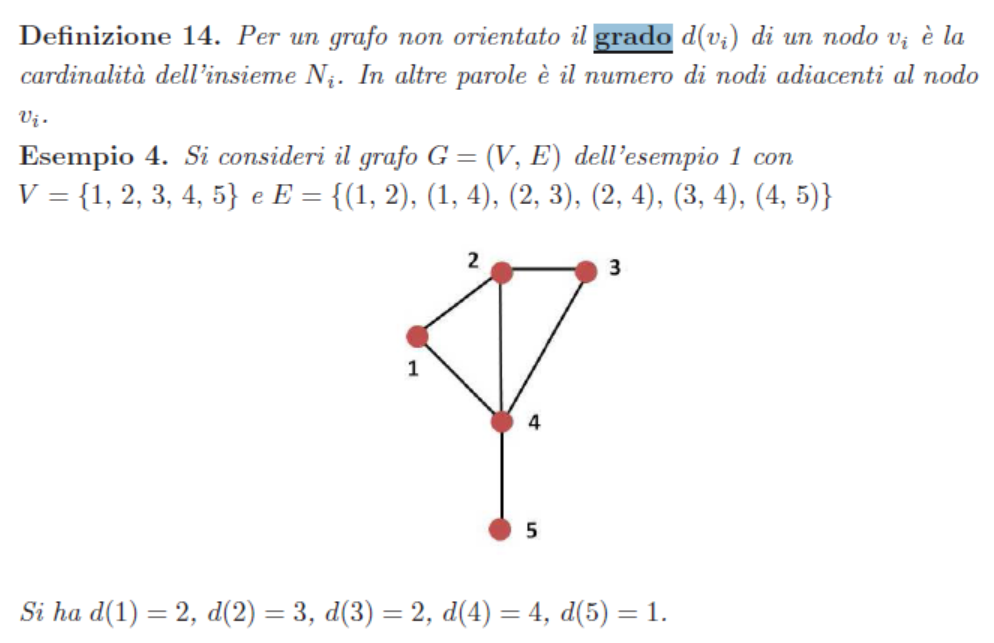
## Teora dei grafi (Teoria delle reti)

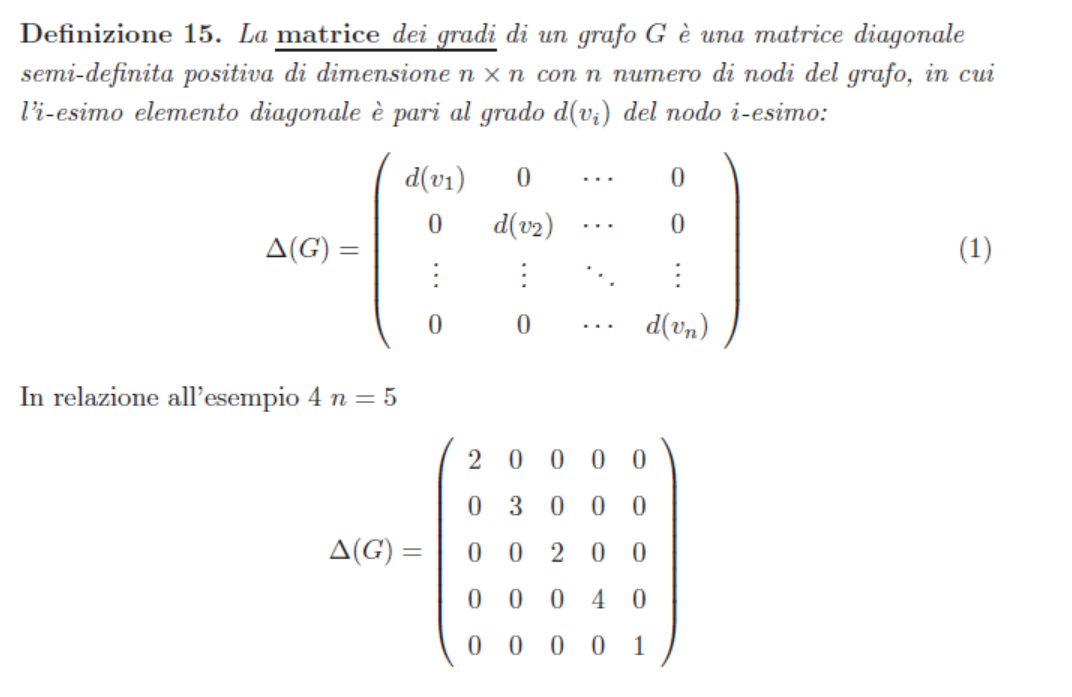
Un *grafo* o una *rete*, sono sinonimi, è una collezione di **nodi** (detti nodes o legami) ed **archi** (edges o legami). Tipicamente un grafo in matematica è definito da una matrice matematica, chiamata ***matrice di adiacenza***. La matrice di adiacenza è una matrice *nxn* che conterrà nell’ *i-esima* e *j-esima* posizione 1 se c’è un legame (cioè se c’è un arco) tra l’elemento i-esimo e j-esimo, altrimenti 0 se non c’è.

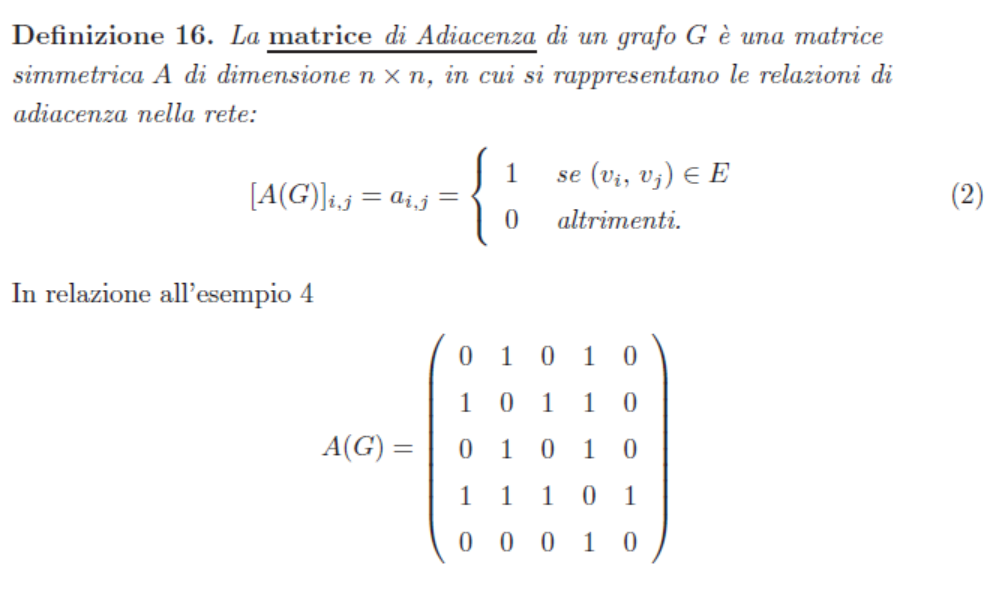
****

### Nodi adiacenti



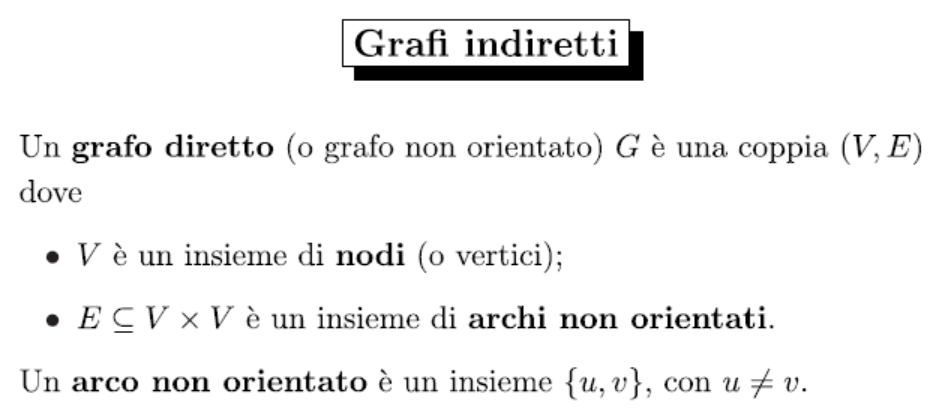


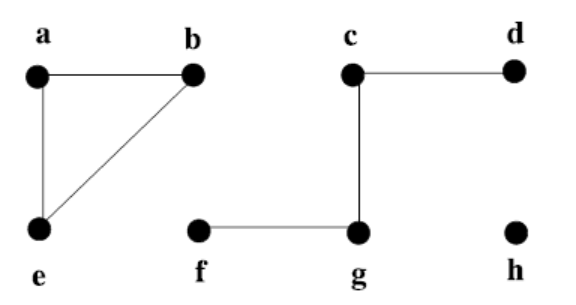




### Grafi indiretti

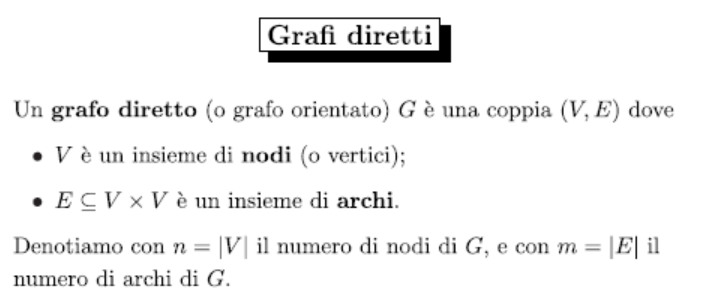
I grafi indiretti sono l’esempio più semplice di grafi. I cerchi sono i nodi e gli archi sono le linee. Di seguito c’è la corrispondenza matrice di adiacenza, che qui è booleana. La matrice di adiacenza di un grafo indiretto è simmetrica, a causa della mancanza di direzione degli archi. Come ad esempio l’amicizia su Facebook che è simmetrica.

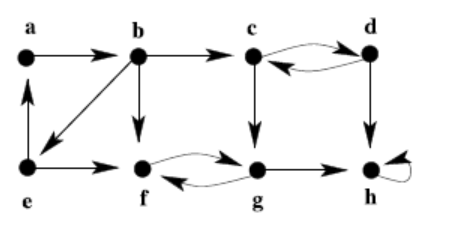


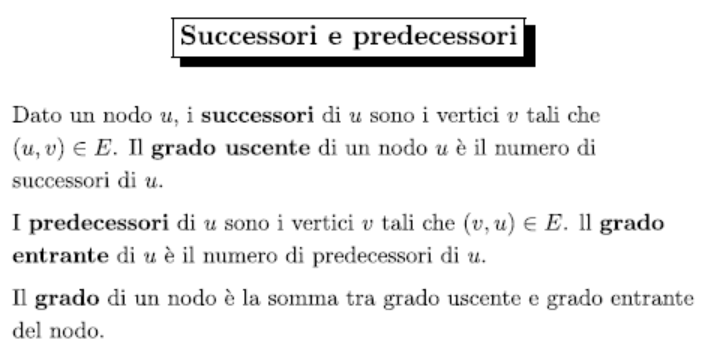


### Grafo diretto

Ad esempio, il follow su Twitter. A differenza dei grafi diretti, nei grafi indiretti ci può essere un **cappio**, cioè una relazione che parte da un nodo ed arriva allo stesso nodo. La corrispondente matrice di adiacenza non è simmetrica. Sulla diagonale stanno le relazioni che partono e arrivano sullo stesso elemento







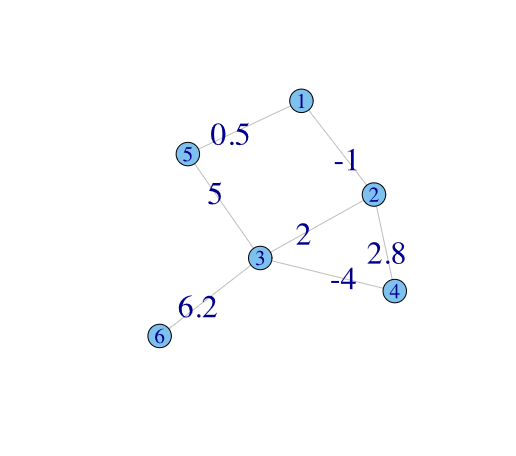
### Grafi semplici e grafi multipli

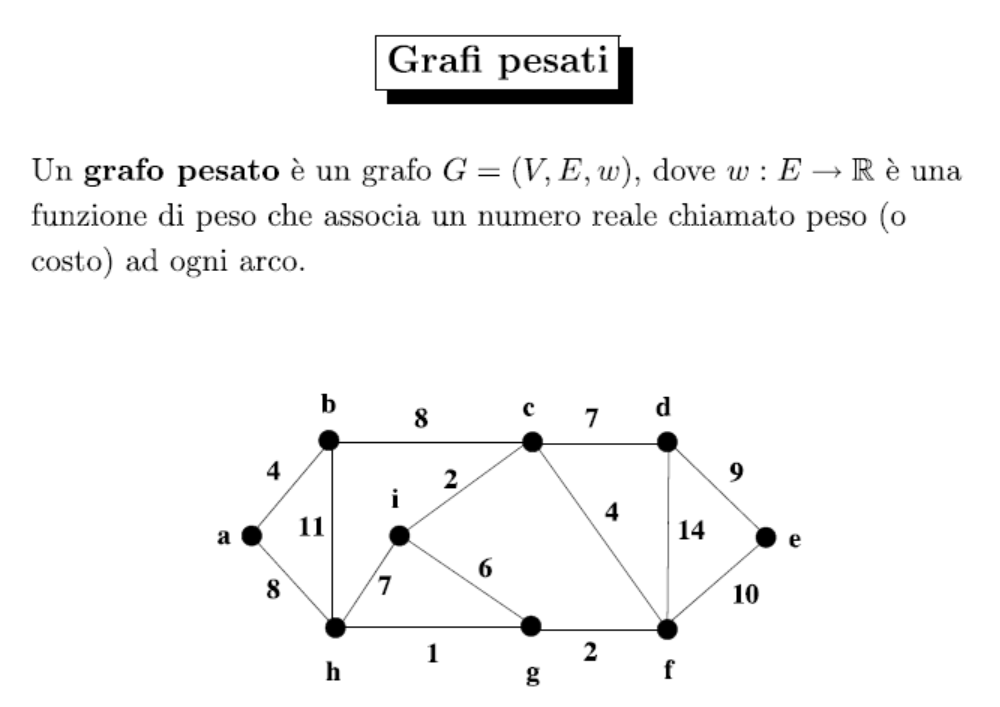
Grafi semplici: se c’è al più un solo arco fra 2 nodi. Sono quelli visti finora.

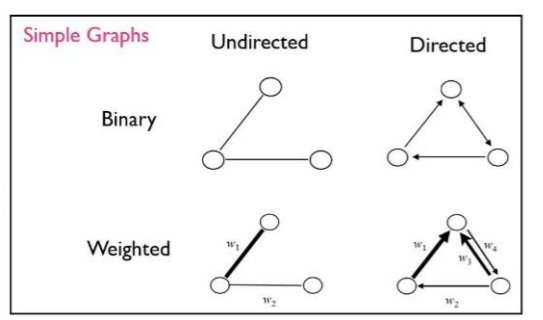
Grafi multipli o **multigrafi**: I grafi multipli sono quelli in cui ci sono molti archi fra coppie di nodi. Ad esempio, le strade fra 2 città, ci possono essere più strade che portano da una città all’altra. In forma tabellare i grafi multipli vengono rappresentati non limitando la cella ai due valori 1 e 0, ma si mette il numero di archi fra i nodi. Ciò ricorda molto i grafi pesati.

### Grafi pesati

I grafi pesati sono grafi in cui su ogni arco è definito un numero reale (spesso, ma non sempre) positivo detto **peso**. Ad esempio, il peso potrebbe indicare la quantità di informazione che passa mediamente per un router, oppure la durata dell’amicizia fra due persone. Un esempio di peso negativo è quello nelle reti sociali in cui un soggetto può avere un certo grado di amicizia (positivo) o un certo grado di inimicizia (negativo). La rappresentazione avviene tramite matrice di adiacenza definita con celle che hanno come numero il peso o 0. Attenzione che un arco di peso 0 corrisponde all’assenza di relazione.







(salto a Paths and cycles)

### Cammino

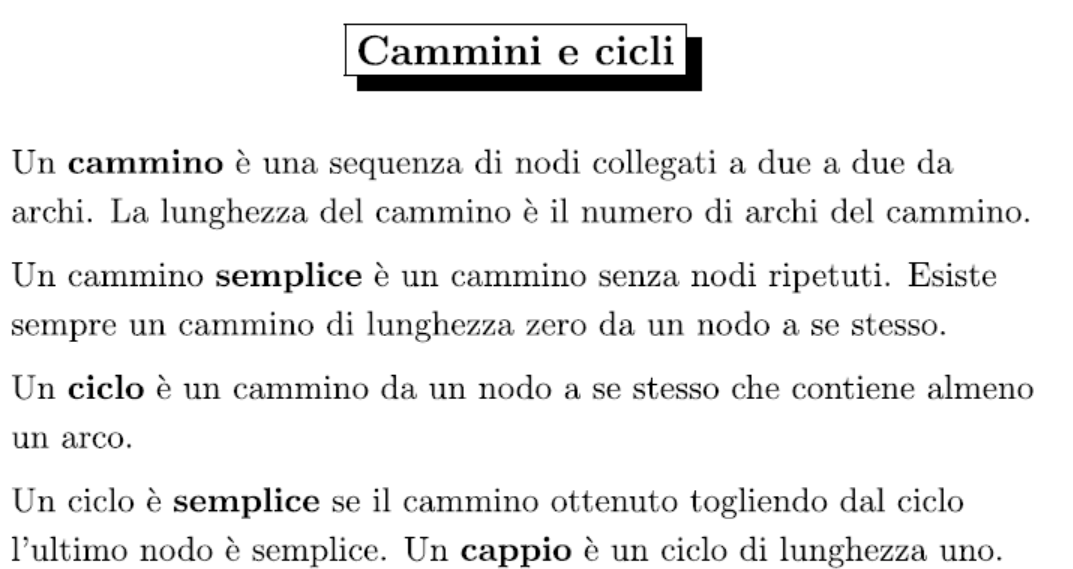
Un **cammino** su un grafo è una sequenza di nodi connessi da archi. cammino è una sequenza di archi che connettono una sequenza di nodi.

un percorso è un cammino in cui un nodo appare al massimo una volta nella sequenza (controllare).

*Cammino non semplice* è un cammino con nodi ripetuti. *Cammino semplice*: cammino con nodi non ripetuti. Ad esempio, (1,4,4,6) è un cammino non semplice.

### Ciclo

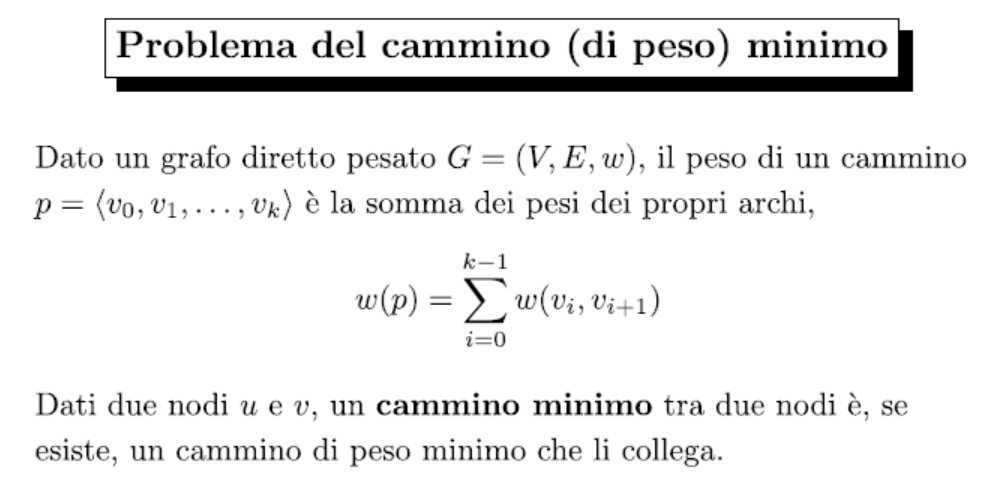
Un **ciclo** è un cammino con nodo partenza = nodo arrivo. Ad esempio, (2,5,3,2) che è un cammino diverso da (5,3,2,5). L’ordine è importante nelle sequenze, e quindi nei cammini e nei cicli.



### Cammino minimo

- Cammino minimo su un grafo non pesato è un cammino con il minor numero di archi tra 2 nodi.

- Cammino minimo su un grafo pesato è cammino con minima somma dei pesi sugli archi del cammino.



### Lunghezza

In un grafo non pesato la lunghezza di un cammino è il numero di archi.

In un grafo pesato la lunghezza del cammino è la somma dei pesi sugli archi del cammino, quindi si dirà cammino minimo se il cammino è di peso minimo. Non è detto che un cammino minimo esista. E nemmeno che sia unico.

### Grafi connessi e fortemente connessi

Un grafo indiretto è **connesso** se esiste un’unica **componente** connessa, ossia se per ogni coppia di nodi esiste un cammino che li unisce (detta altrimenti, se ogni coppia di nodi è raggiungibile da un cammino).

Un grafo diretto si dice **fortemente connesso** (ocio) se per ogni coppia di nodi (i,j) esiste un cammino da i a j ed un altro cammino da j a i. Per esempio, un grafo diretto ciclico.

### Componenti connesse e fortemente connesse

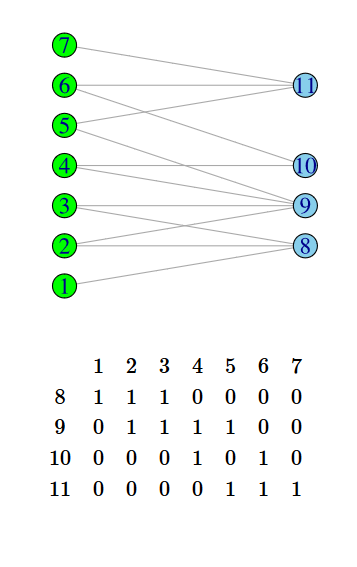
Si chiamano **componenti connesse** il massimo insieme dei nodi che sono connessi in un grafo. Detto altrimenti, sono i sottoinsiemi massimali di nodi che sono connessi nel grafo. I sottoinsiemi minimali che sono fortemente connessi sono detti **componenti fortemente connessi**. Un grafo è connesso se ha un’unica componente fortemente connessa.

Salto Bipartite graphs and projections

### Grafo bipartito

I grafi bipartiti sono dei grafi per i quali si possono suddividere i nodi in due gruppi, tali che gli archi vanno solo da nodi di un tipo a quelli dell’altro. Ad esempio, attori a sinistra e film a destra.

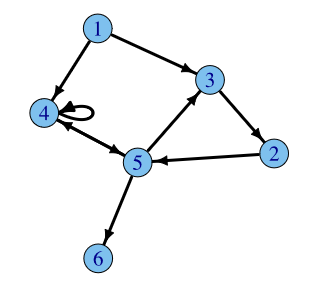
I grafi bipartiti sono degli esempi di grafi indiretti, per i quali solitamente si usa un’altra matrice che è chiamata ***matrice di incidenza***: se i due gruppi hanno cardinali n e k, la matrice avrà ampiezza *nxk* In questo modo le righe saranno gli elementi di un gruppo e le colonne dell’altro, e si usa la medesima idea di porre 1 o 0 nelle celle se esiste o meno la relazione tra due elementi.



### Grado di un nodo

Il grado di un nodo nel caso di grafo indiretto è il numero di **nodi vicini**, cioè quelli che sono raggiungibili con un arco (cioè in un solo passo). Ad esempio, in figura a lato il grado del nodo 3 è 4. Il grado di un nodo è ottenibile sommando la corrispondente riga della matrice di adiacenza (oppure la colonna essendo simmetrica).

La nozione si triplica nel caso di grafi diretti:

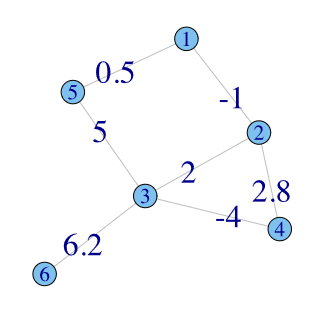
**- Grado uscente**: il numero di nodi (ocio) che posso raggiungere da un nodo. Numero dei successori. Nella tabella si somma la **riga**, sulla riga vedo i nodi che raggiungo. In figura a fianco: grado uscente di 4 è 2.

- **Grado entrante**: il numero di archi (ocio) che mi portano al nodo. Numero dei predecessori. Nella tabella si somma la **colonna**, sulla colonna vedo i nodi che mi raggiungono. In figura a fianco: grado entrante di 4 è 2.

**- Grado totale**: la somma di grado uscente ed entrante.

Nb. Il cappio è sia grado entrante che grado uscente.

### Grado pesato di un nodo

Dato un grafo pesato, il **grado pesato** di un nodo è somma dei pesi degli archi che collegano quel nodo. In figura a lato: il grado pesato di 1 è -0.5.

Nel caso di grafi diretti avremo un grado uscente pesato e un grado entrante pesato.

30 \4\ 2018

Reg 326

Arriviamo così alle reti (tutto in 4 ore circa)

Visione Ted talks Manuel Lima

Abbiamo visto fino ad ora dati che possono essere archiviati in forma tabellare, ma ci sono dati che “sfuggono” a questa griglia: testo, dati semi-strutturati e grafi/reti.

<http://users.dimi.uniud.it/~massimo.franceschet/ds/r4ds/syllabus/make/madrid/madrid.html>

**Network Science**

* learn [Network Science in R - A Tidy Approach](https://www.datacamp.com/courses/network-science-in-r-a-tidy-approach)
* The igraph package
  + glance [igraph](http://igraph.org/)
  + learn Getting started with igraph [html](http://users.dimi.uniud.it/~massimo.franceschet/ns/syllabus/make/igraph/igraph.html) + [Rmd](http://users.dimi.uniud.it/~massimo.franceschet/ns/syllabus/make/igraph/igraph.Rmd)
* The ggraph package
  + glance [ggraph](http://www.data-imaginist.com/2017/Announcing-ggraph/)
  + learn Getting started with ggraph [html](http://users.dimi.uniud.it/~massimo.franceschet/ns/syllabus/make/ggraph/ggraph.html) + [Rmd](http://users.dimi.uniud.it/~massimo.franceschet/ns/syllabus/make/ggraph/ggraph.Rmd)
* The tidygraph package
  + glance [tidygraph](https://cran.r-project.org/package=tidygraph)
  + learn [Getting started with tidygraph](https://www.data-imaginist.com/2017/introducing-tidygraph/)

Nel mondo reale possiamo modellare del modo reale in 5 categorie

1. Le reti tecnologiche: internet

2. Le reti sociali

4. Reti di informazione: come il web

5. Reti di citazioni fra articoli scientifici o brevetti

6. Reti biologiche: interazione proteina-proteina, reti alimentari

Si modellano i dati come insieme di punti collegati fra di loro. Quello che conta è la relazione fra i nodi, il punto senza la relazione non è molto significativo.

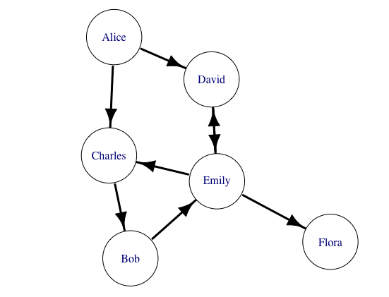
Tramite il pacchetto ***igraph*** di R analizziamo le reti. E’ il corrispettivo di dplyr nel senso che così come dplyr permette di analizzare dataframe, igraph permette di analizzare reti e quindi calcolare misure di centralità, cammini minimi, calcolare gruppi di comunità, distribuzione del grado dei nodi,etc.

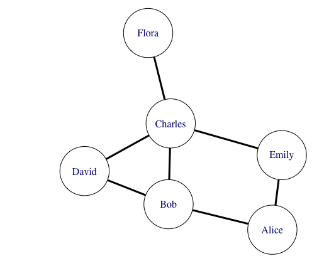
Tramite il pacchetto ***ggraph*** di r visualizziamo le reti. E’ il corrispettivo di ggplot. Aggiungiamo delle geometrie per archi e nodi e visualizzare le reti.

Non vedremo il pacchetto *tidygraph*, perché siamo solo all’inizio.

# Network science in R - A tidy approach

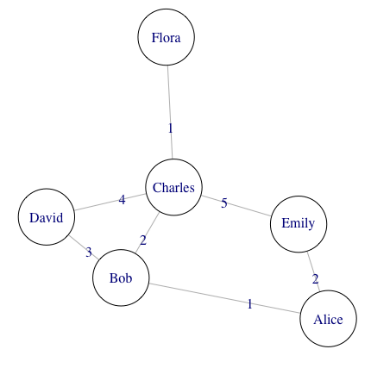
Le entità sono i nodi, e le relazioni sono gli archi. Eulero ha cominciato ad usarli (7 ponti di Colinsberg)

**Rete diretta**

Se hanno le frecce. Sono le più semplici.

**Rete indiretta**

Gli archi non hanno direzione

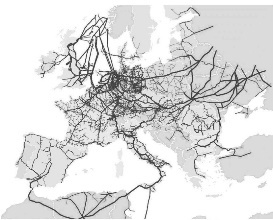
**Reti pesate e non pesate**

Anche le reti indirette possono essere pesate. Nelle reti pesate gli archi hanno associato un numero reale che ne quantifica la forza della relazione, se è negativo allora la relazione è inversa. Possiamo anche avere pesi nulli, ovvero assenza di relazione.

Definiremo una metrica per misurare la relazione: ad esempio la quantità di bit scambiati fra server, quantità di commenti ai post di un altro. I pesi nulli significano l’assenza di relazione (come indice di correlazione in statistica).

L’unica caratteristica fondamentale che tutte le reti hanno in comune è che sono *webs without a spider*, non c’è un’unità centrale che coordina e controlla la rete. Il tutto viene in modo organizzato autonomamente, gli agenti si auto-organizzano. Ad esempio, posso costruire un sito web senza chiedere permesso a Roma, il web è decentralizzato e autorganizzato. Un altro esempio è lo stormo di uccelli che con 3 regole fa si che si formino le figure che vediamo nei cieli. Le regole sono: non sovrapporsi ai simili, non allontanarsi troppe e se i vicini virano vira anche lui.

La numerosità degli individui assieme al comportamento molto semplice degli agenti può creare una *proprietà emergente* visibile a livello globale. C’è un dunque un passaggio da micro a macro, interessante nelle reti complesse.



La network science può essere considerata come uno spin-off della data science. L’obiettivo è quello di analizzare e visualizzare i dati di rete. Un esempio di output è la rete di gasodotti europea.

**3 rappresentazioni in corrispondenza**

Una rete ha tre rappresentazioni tutte in corrispondenza:

* grafo
* matrice
* dataframe

Posso passare da Matrici a grafo, da grafo a dataframe, da daframe a matrice. Le tr diverse rappresentazioni sono comode in diversi problemi: per calcolare lo spettro del grafo è comoda la matrice del grafo, per calcolare una misura di centralità (pagerank) è meglio l’oggetto di igraph, per visualizzare il grafo con gpraph o fare un raggruppamento con dplyr è meglio la rappresentazione a DFrame.

Agenda:

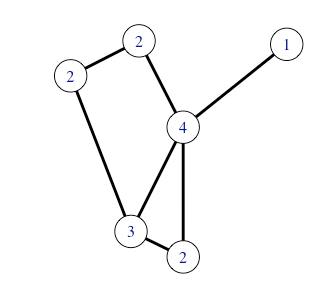
* Capitolo 1: **misura di centralità dei nodi**: vedremo 2 modi per determinare quali sono i nodi importanti di una rete
* Capitolo 2: **misura di centralità per gli archi**: vedremo quando le connessioni sono importanti. Vedremo la *teoria della forza dei legami deboli*. Ne vedremo una.
* Capitolo 3: **misura di similarità fra coppie di nodi**. Vedremo dei modi per stabilire quando 2 nodi sono simili, a seconda del pattern che hanno. Ne vedremo una.
* Capitolo 4: metodo per trovare gruppi di nodi simili basati sul **clustering gerarchico**

## 1) Misure centralità dei nodi

È una funzione che assegna un rating, un numero che denota l’importanza, ad ogni nodo. Avendo ogni nodo della rete un punteggio otteniamo un ranking, **una** classifica di importanza. Con centralità si intende l’importanza di ogni nodo.

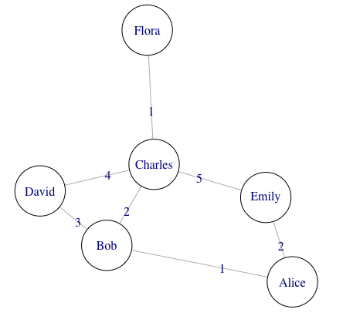
Ad esempio, Google usa un algoritmo chiamato *pagerank*, che è una misura di centralità ricorsiva, per ottenere i risultati in un certo ordine, per prime le pagine con centralità maggiore. Un altro esempio è la valutazione della ricerca dei **ricercatori** in termini di citazioni sui loro articoli. Ancora un altro esempio, su internet potremmo cercare di capire quali sono le macchine più importanti per salvaguardarle da un attacco terroristico o da una rottura di tubi.

La nozione di centralità non è univoca. Ogni misura da l’importanza a un determinato aspetto della rete, è bene non fissarsi su una singola misura ma considerarne tante. Ne vedremo solo 2.



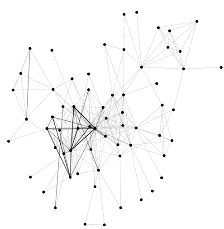
**Grado di un nodo** (1 di 4)

La prima misura di centralità è il grado di un nodo. Un nodo è importante se ha molte connessioni. Ossia se ha molti archi. È una misura semplice e abbastanza affidabile, ad esempio nelle reti di citazioni si usa per contare le citazioni. Nel caso di reti dirette abbiamo due misure: grado uscente e grado entrante. Ad esempio, in figura il nodo 4 ha grado 4, il nodo 2 ha grado 2.

Nel caso di reti dirette abbiamo due misure: **grado uscente** e **grado entrante**. Dobbiamo quindi specificare il tipo di grado, anche se in realtà conta molto di più il nodo entrante, visto che è meno controllabile (giudizio della comunità è più importante rispetto a quello che uno può dare alla comunità).

**Grado pesato (o forza)**

Solo nelle reti pesate c’è un’altra misura di centralità chiamata grado pesato. Non si conta il numero di archi, ma si somma il peso degli archi. È la somma dei pesi degli archi incidenti il nodo in esame.

Come esempio di studio useremo la rete terroristica che ha organizzato gli attacchi a Madrid nel 2004. Ogni arco è etichettato con un numero intero da 1 a 4 a seconda del di quante relazioni sono soddisfatte (4 al massimo). Challenges 7: quali sono stati i terroristii più importanti nell'esplosione del treno di Madrid del 2011? [file html](http://users.dimi.uniud.it/~massimo.franceschet/ds/r4ds/syllabus/make/madrid/madrid.html) e [file Rmd](http://users.dimi.uniud.it/~massimo.franceschet/ds/r4ds/syllabus/make/madrid/madrid.Rmd).

Usiamo i pacchetti dplyr, readr ggplot2, igraph ggraph, visNetwork.

g <- graph\_from\_data\_frame(edges, directed = FALSE, vertices = nodes)

print(g)

IGRAPH c629d58 UNW- 64 243 --

+ attr: name (v/c), weight (e/n)

+ edges from c629d58 (vertex names):

‘UNW’ è acronimo di Undirected (indiretto) Weighted (pesato). Il grafo ha 64 e 243 archi. ‘v’ = vertici (o nodi) ‘e’= edges (o archi). ‘name ‘ è attributo di vertici di tipo ‘c’ carattere. ‘weight’ è un attributo di edges ti tipo ‘n’ numeric.

* V(g) restituisce l'insieme dei nodi del grafo --> lista terroristi
* vcount(g) restituire il numero dei nodi
* V(g) analogamente per gli archi
* ecount(g) analogamente per gli archi

Le proprietà sono attributi che posso assegnare al grafo, o ai singoli nodi e ai singoli archi.

* g$name <- "Madrid network" #assegno un nome all’attributo name del grado. “nome del grafo $ attributo <- nome da assegnare".
* V(g)$id <- 1:vcount(g) # assegno i numeri da 1 a 64
* E(g)$weight # vedo i pesi

Alcune visualizzazioni

Il grafo è unico, visualizzabile in molti modi.

* set\_graph\_style() # imposta lo stile a quello di grafo: rimuove gli inutili assi cartesiani
* ggraph(g, layout = "with\_kk") + geom\_edge\_link(aes(alpha = weight)) + geom\_node\_point() # imposto layout basato su algoritmo Kamada-Kawai tale che nodi con più archi saranno più vicini nel grafico. C’è poi una geometrica che disegna gli archi, qui in proporzione al peso (più scuri). L’ultima geometria disegna i nodi come punti
* + geom\_node\_text(aes(label = id), repel=TRUE) # aggiunge l’etichetta id al nodo. Repel = T non fa sovrapporre le etichette
* ggraph(g, layout = "in\_circle" # dispone in cerchio i nodi
* ggraph(g, layout = "grid " # dispone i nodi sulla griglia

Registrazione 327

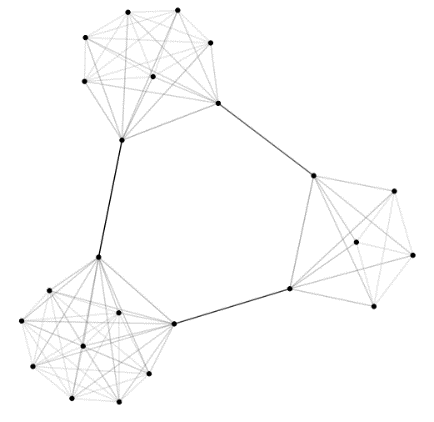
**Closness**

Misura quanto un nodo è vicino agli altri nodi. La closness di un nodo è il reciproco della distanza media di un nodo dagli altri. Dove con distanza media si intende la distanza geodetica, ossia il numero di passi che deve fare un nodo per raggiungere un altro nodo (lunghezza del cammino minimo).

Registrazione 328

## 2) Misura centralità per archi

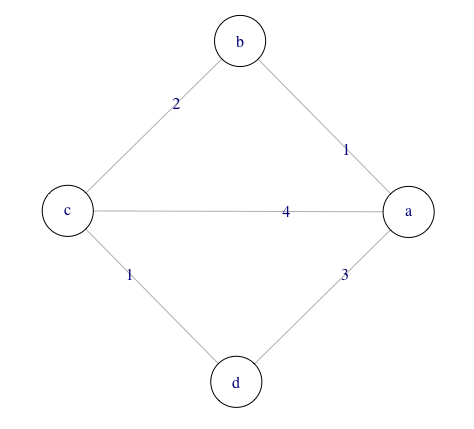
**Betweenness**

****È una misura di centralità che useremo sugli archi, ma che si può usare anche seui nodi. Misura il numero di cammini di altri nodi passano sull’arco. Tanti più cammini passano, tanto più importante sarà quell’arco. Tanta più forza avrà.

In figura: ci sono 3 archi (grassetto) che connetto le 3 comunità di nodi. Ogni cammino che connette due comunità deve passare per i 3 archi. Essi avendo tanti cammini in un certo senso controllano l’informazione. Se li rimuovessimo le 3 comunità sarebbero isolate.

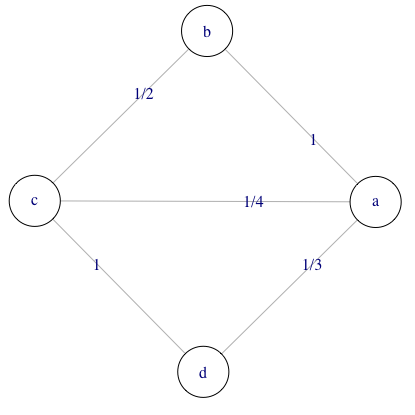
È il numero di cammini minimi, che connettono un’altra coppia di nodi, che passano per un fissato arco. Conta quante volte un arco è attraversato da cammini minimi. Se ho ben capito: fissata una coppia di nodi considero la frazione di cammini minimi che passano su un arco fissato rispetto alla frazione di cammini minimi che collegano I due nodi.

Lo stesso concetto vale anche per I nodi. Invece di selezionare il numero di cammini minimi che passano per un arco, seleziono il numero di cammini minimi che passano per un nodo.

**Betweenness in reti pesate**

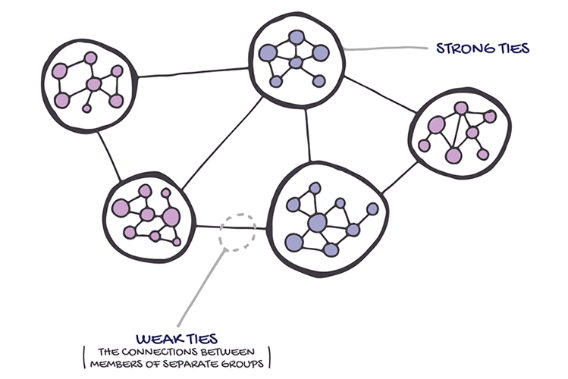
Il peso di un cammino è la sommatoria dei pesi degli archi del cammino. Il cammino minimo fra due nodi è il cammino con peso minimo. Ad esempio, in figura a fianco il cammino minimo fra a e c è (a,b,c) con peso 3. Nota che non ha il numero minimo di archi e ciò dà problemi di incoerenza. Perciò vd poi.

Prima di calcolare la betweenness in archi pesati dobbiamo invertire il peso degli archi. Così archi con peso basso sono preferiti rispetto ad archi con peso alto, nell’ottica di una minimizzazione\ Betweenness.

(index1)

In sintesi: dopo l’inversione dei pesi, archi con peso basso corrispondono a nodi molto vicini e che sono preferiti rispetto ad archi di peso alto che corrispondono a nodi distanti.

La misura di centralità appena introdotta permette di verificare la teoria sociologica Teoria sociologica de *la forza dei legami deboli* di Martin Granovetter (anni ‘70).

Ties = archi=legami.

Granovetter dice che se prendiamo una rete sociale si formano delle comunità molto coese connesse da archi sporadici.

Gli archi dentro le comunità sono chiamati archi forti (*strong ties*). Gli archi fra le comunità sono chiamati archi deboli (*weak ties*), che sono relazioni fra membri di comunità diverse.

Da un punto di vista sociologico un **legame debole** è una relazione fra membri di comunità diverse. I legami deboli non hanno bisogno di grosso accudimento per rimanere vive. Ad esempio, un amico di penna americano. Un **legame forte** è una relazione fra persone che condividono tante esperienze, e necessitano per rimanere attive di molta cura. Ad esempio, i legami fra amici intimi o compagni di studi. Questi due legami hanno un’importanza molto differente. La ricerca dimostrò che i legami forti tendono a generare idee dominanti e stagnanti, viceversa i legami deboli tendono a favorire la diversità delle idee e un modo diverso di pensare. Il sociologo, con molta sorpresa, dimostrò in un articolo che i legami più importanti sono i legami deboli, in quanto danno la possibilità di accedere a mondi nuovi e modi di pensare diversi da quelli della propria comunità. Un esempio è dato dal fatto che l’amico di penna ha un intero mondo: di amici, relazioni, fidanzata, lavoro.

Ma la teoria vale anche per la rete terroristica del nostro esempio? Definiamo legame debole quel legame con un peso pari a 1. Verifichiamo se i legami che hanno una connessine peso pari a 1 sono importanti o frequenti rispetto agli altri legami.

Continua da registrazione 329

Vai su Rmd

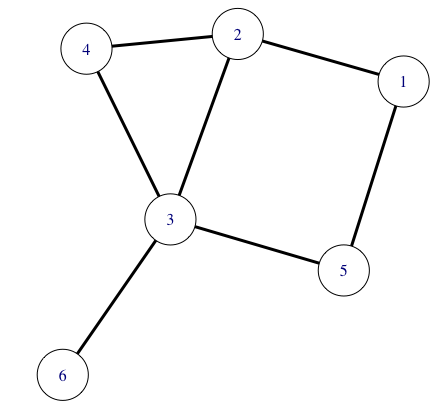
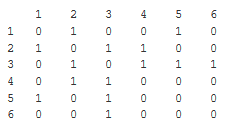
Calcoliamo la betweenness

* dist\_weight <- 1 / E(g)$weight # imposta l
* btw <- edge\_betweenness(g, weights = dist\_weight) # con la funzione edge\_betweenness calcoliamo la betweenness, pesata attraverso weights. Lo fa solo sui cammini minimi
* ggraph(g, layout = "with\_kk") + geom\_edge\_link(aes(filter = weakness), alpha = 0.5) +
* geom\_node\_point() # posso usare l’estetica filter per filtrare i legami deboli, infatti weakness è booleana T se il legame è debole.

## 3) Misura di similarità fra coppie di nodi

Quando due nodo sono simili fra loro? Lo scopo è capire se 2 nodi sono simili. Trovare nodi simili è un problema interessante, ad esempio è utile trovare pagine web simili, Amazon cerca spesso clienti simili. Il problema della similarità è un problema binario, abbiamo due nodi e vogliamo capire quanto siano simili fra di loro. Diremo che due nodi sono simili se il pattern di connessione è simile. Dati due nodi S e T e un terzo nodo X, possiamo fare considerazioni di questo genere: se sono entrambi in relazione con X allora c’è più similarità, se S è in relazione con X e T no allora la similarità sarà maggiore.

Più sono simili le righe della matrice di adiacenza più simili sono i nodi (di riga). La similarità può essere misurata tramite il coefﬁciente di correlazione di Pearson [-1,1]. Come



Diremo che 2 nodi sono simili se essi hanno lo stesso pattern di connessione di nodi vicini. Guardo alle righe per vedere se 2 nodi sono simili. Ma non sommo! Faccio indice di pearson. Diremo che 2 nodi sono simili se hanno coeff di correlazione alto, (in posizione 1,1 = 1,2 etc) vicino ad uno. Sono invarianti se coefficiente tende a zero.

* A <- as\_adjacency\_matrix(g, attr = "weight", sparse = FALSE, names = FALSE) # passaggio da grafo a matrice di adiacenza: uso la funzione as\_adjacency\_matrix che come oggetto vuole un grafo, attributo il pes, sparse dice se voglio una matrice sparsa
* rowSums(A) # fa somma righe, quindi calcola la forza\ grado pesato.
* B = A > 0, rowSums(B) # se non voglio il grado trasformo la matrice in booleani e poi sommo le righe, ottenendo il grado semplice
* S <- cor(A), diag(S) = 0 # colcolate le correlazioni metto a 0 la diagonale (è 1)

Poiché la massima similarità è raggiunta da nodi con grado molto basso, allora filtro gli archi al di sora di un grado minimo

* We # (dovrei scrivere altri comandi, visto che li chiede nel compito scritto)

La matrice di adiacenza non è simmetrica per gradi diretti, lo è per grafi indiretti.

La matrice di adiacenza per grafi pesati riporta il numero del peso.

Vai su datacamp

Registrazione 330

(ripasso lungo per assenti)

07/05/2018

registrazione 337

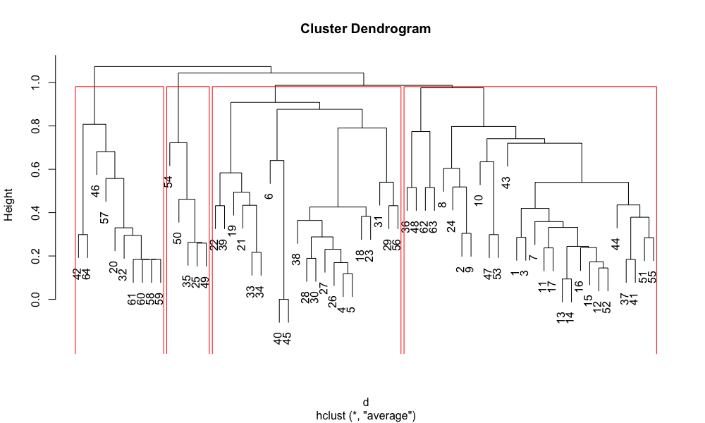
Misure di similarità:

* Correlazione fra coppie: misura correlazione fra coppie di riga e colonna (eliminare?)

## 4) Clustering gerarchico

Metodo che utilizza come distanza fra due nodi l’inverso della similarità. Infatti, se 2 nodi sono simili hanno una similarità elevata, allora la loro distanza è molto bassa. E viceversa. Come misura di similarità usiamo quella calcolata al punto 3.

Faccio tante classi quanti sono i nodi, poi ad ogni passo unisco due gruppi alla volta. Calcolo la distanza minima fra tutti i nodi, che diventa il criterio per unire i gruppi in uno step. Unisco i gruppi la cui distanza è ?maggiore? della distanza minima. La distanza minima cambierà aumentando, ad ogni turno. Avrò tante soluzioni. Otterremo così un *dendrogramma* dove i nodi dell’albero (ocio) sono le unioni dei gruppi e le foglie sono i nodi del grafo iniziale. Non sapendo quanti gruppi è meglio avere, devo decidere a che altezza fra 0 e 1. potare l’albero (cut): taglio tracciando linea orizzontale, ogni volta che incrocia un alinea verticale è determinato un gruppo. Nell’esempio sono 4 gruppi. L’analista scelta la soluzione intermedia fra n e 1 gruppi.

La distanza è il complemento a 1 della similarità.

C’è un ultimo problema da affrontare. Non sappiamo cosa sia la similarità fra gruppi.

**Similarità fra gruppi**

* **Single-linkage**: la similarità fra tra due gruppi è definita come il massimo delle similarità tra i nodi di diversi gruppi.
* **Complete-linkage**: similarità fra gruppi è definita come il minimo delle similarità fra i gruppi di nodi
* **Average-linkage**: dove la similarità tra due gruppi è la media delle similarità tra i nodi di diversi gruppi (più frequente)

Il metodo di clustering gerarchico è il seguente:

* Valutare le misure di similarità per tutte le coppie di nodi
* Assegnare ciascun nodo a un gruppo a sé stante.
* Trova la coppia di gruppi con la similarità più elevata (ossia distanza minore) e uniscili in un unico gruppo.
* Calcola la somiglianza tra il nuovo gruppo composito e tutti gli altri.
* Ripetere i passaggi 3 e 4 fino a quando tutti i nodi sono stati uniti in un singolo gruppo.

L’algoritmo non sveglie il numero k di gruppi, è scelto dall’analista.

* D <- 1-S, d <- as.dist(D) # calcolo matrice di distanza come complemento a 1 della similarità
* cc <- hclust(d, method = "average"), plot(cc) # hclust calcola il clustering gerarchico e visualizzo
* cls <- cutree(cc, k = 4) # taglio il dendrogramma t.c. formi 4 gruppi
* nodes <- mutate(nodes, cluster = cls) # aggiungo poi tale classificazione al DFframe, che conterrà per ogni nodo la sua appartenza ad uno dei 4 gruppi.
* filter(nodes, cluster == 1) %>% select(name)# posso filtrare al solito con dplyr i nodi
* V(g)$cluster <- nodes$cluster # aggiungo informazione alla rete attraverso l’attributo cluster dei nodi
* ggraph(g, layout = "with\_kk") +

geom\_edge\_link(aes(alpha = weight), show.legend=FALSE) +

geom\_node\_point(aes(color = factor(cluster))) +

labs(color = "cluster") # nella geometria geom\_node\_point indico il colore

## Visualizzazione interattiva delle reti con Visnetwork

Permette di zoomare, selezionare e spostare i nodi: che fungono da elastico, c’è una forza di repulsioni sui nodi e di attrazione sugli archi.

VisNetwork è un pacchetto eccellente per realizzare bellissime visualizzazioni interattive delle reti. Posso visualizzare una rete personalizzata o una rete creata con il pacchetto igraph. Posso anche dare alla rete diversi layout. Ancora più importante, le reti che disegni sono vivi! Ciò significa che puoi interagire con i nodi e i bordi o con l'intero grafico. Ad esempio, puoi fare clic e evidenziare un nodo, spostarti intorno a un nodo o spostare e zoomare l'intera rete. Godrai il piacere dell'interazione durante gli esercizi.  
Posso selezionare un nodo tramite il suo identificatore, nel nostro caso il nome dei terroristi. Supponiamo che tu conosca il nome del terrorista e desideri localizzarlo sulla rete, insieme ai suoi vicini. È   
 possibile suddividere i nodi in gruppi, ad esempio, in cluster di nodi simili, come abbiamo fatto nella nostra applicazione, e evidenziare tutti i nodi in un particolare gruppo.

# Esercizio in classe sui delfini

Reg 339

<http://users.dimi.uniud.it/~massimo.franceschet/ds/r4ds/syllabus/make/dolphin/dolphin.html>

David Lusseau, ricercatore presso l'Università di Aberdeen, osservò il gruppo di delfini di Doubtful Sound. Ogni volta che una scuola di delfini si incontrava nel fiordo tra il 1995 e il 2001, ogni membro adulto della scuola veniva fotografato e identificato da segni naturali sulla pinna dorsale. Questa informazione è stata utilizzata per determinare quanto spesso due individui sono stati visti insieme. Leggi la storia completa.

download

[Csv con nome e sesso delfini](http://users.dimi.uniud.it/~massimo.franceschet/ds/r4ds/syllabus/make/dolphin/dolphin_nodes.csv)

[Csv con legami fra delfini](http://users.dimi.uniud.it/~massimo.franceschet/ds/r4ds/syllabus/make/dolphin/dolphin_edges.csv)

Domande:

* Quali sono i delfini più socievoli?
* I delfini femmine sono più socievoli dei delfini maschi?
* Qual è il tipo di rapporto tra i delfini (amicizia o sessuale)?

Reg 340

Soluzione:

* Quali sono i delfini più sociali? Bastava calcolare il grado: sono Grin sn4, Topless, scabs, trigger.
* I delfini femmine sono più socievoli dei delfini maschi? C’è una differenza ma non significativa
* Qual è il tipo di rapporto tra i delfini (amicizia o sessuale)? Più amicale che sessuale.

Comandi soluzione

* r # (forse dovrei inserirne alcuni)

Assortatitività: cerchio = femmina, quadrato = maschio L R

1 cerchio -- quadrato 3

1 cerchio --- quadrato 2

poi

usare correlazione, oppure contare i rapporti MM + FF vs MF

Comodo usare vettori booleani:

SexU, sexF, sexM

**Biological Networks**

<http://users.dimi.uniud.it/~massimo.franceschet/teaching/datascience/network/biological.html>

Le reti sono ampiamente utilizzate in molti settori della biologia come una comoda rappresentazione di modelli di interazione tra elementi biologici appropriati. Queste reti biologiche includono reti biochimiche, reti neurali e reti ecologiche.

## Reti biochimiche

**Le reti biochimiche** rappresentano i modelli a livello molecolare dell'interazione e i meccanismi di controllo nella cellula biologica. I principali tipi di queste reti sono reti metaboliche, reti proteiche-proteiche e reti di regolazione genetica.

### Reti metaboliche

**Il metabolismo** è il processo chimico mediante il quale le cellule suddividono cibo e sostanze nutritive in blocchi utilizzabili e quindi riuniscono questi elementi costitutivi per formare le molecole biologiche che la cellula ha bisogno per completare gli altri compiti. Tipicamente questa suddivisione e riassemblamento comporta catene o **percorsi** , serie di reazioni chimiche successive che convertono gli input iniziali in prodotti finali utili mediante una serie di passaggi. L'insieme completo di tutte le reazioni in tutti i percorsi forma la **rete metabolica** .

I vertici in una rete metabolica sono sostanze chimiche prodotte e consumate dalle reazioni. Queste sostanze chimiche sono conosciute come **metaboliti** . Queste sono piccole molecole come carboidrati, lipidi, nonché amminoacidi e nucleotidi. I metaboliti consumati sono chiamati **substrati** della reazione, mentre quelli prodotti sono chiamati i **prodotti** .

La maggior parte delle reazioni metaboliche non si verificano spontaneamente, o lo fanno solo a un tasso molto basso. Per far sì che le reazioni avvengano a un ritmo utilizzabile, la cellula impiega una serie di catalizzatori chimici, chiamati **enzimi** . Gli enzimi non vengono consumati nelle reazioni che catalizzano. Aumentando o diminuendo la concentrazione dell'enzima che catalizza una particolare reazione, la cellula può attivare o disattivare tale reazione o moderarne la velocità. Gli enzimi tendono ad essere altamente specifici per la reazione che catalizzano.

La rappresentazione più corretta di una rete metabolica è come un **grafico bipartito** . I due tipi di vertici rappresentano metaboliti e reazioni metaboliche, con i bordi che uniscono ciascun metabolita alla reazione a cui partecipa. I bordi sono diretti, poiché alcuni metaboliti (i substrati) entrano nella reazione e alcuni (i prodotti) ne escono. Gli enzimi possono essere incorporati aggiungendo una terza classe di vertici per rappresentarli, con spigoli non diretti che li collegano alle reazioni che catalizzano. Il grafico risultante è una **rete tripartita mista (diretta e non orientata)** .

Tuttavia, le rappresentazioni più comuni del **progetto di** rete metabolicala rete bipartita sui vertici del metabolita. In un approccio, i vertici nella rete rappresentano metaboliti e c'è un margine non diretto tra due qualsiasi metaboliti che partecipano alla stessa reazione, sia come substrati che come prodotti. Chiaramente questa proiezione perde gran parte delle informazioni contenute nella rete bipartita. Un altro approccio più informativo è quello di rappresentare la rete come un grafico diretto con un singolo tipo di vertice che rappresenta i metaboliti e un margine diretto da un metabolita all'altro se c'è una reazione in cui il primo metabolita appare come substrato e il secondo come un prodotto . Questa rappresentazione è più ampia in termini di informazioni rispetto a quella non orientata, ma perde comunque l'associazione dei metaboliti con le reazioni.

### Reti di interazione proteine-proteine

Le proteine ​​interagiscono tra loro e con altre biomolecole, sia grandi che piccole, ma le interazioni non sono puramente chimiche. **Le proteine** sono molecole a catena lunga formate dalla concatenazione di una serie di unità di base chiamate **amminoacidi** . Una volta creata, una proteina non rimane in una forma simile a catena, ma si ripiega su se stessa in una forma piegata la cui forma dipende dalla sequenza di amminoacidi. La forma piegata detta l'interazione fisica che può avere con altre molecole. Quindi, la modalità primaria dell'interazione **proteina-proteina** è fisica piuttosto che chimica, le loro complicate forme piegate si incastrano per creare i cosiddetti **complessi proteici** ma senza lo scambio di particelle che definisce le reazioni chimiche.

In una **rete di interazione proteina-proteina** i vertici sono proteine ​​e due vertici sono collegati da un bordo non orientato se la corrispondente proteina interagisce. Tuttavia, questa rappresentazione omette informazioni utili. Le interazioni che coinvolgono tre o più proteine ​​sono rappresentate da più spigoli e non c'è modo di dire dalla rete stessa che tali spigoli rappresentano aspetti della stessa interazione. Questo problema potrebbe essere affrontato adottando una rappresentazione bipartita, con proteine ​​e interazioni come diversi tipi di vertici, e le proteine ​​di connessione dei bordi non orientati alle interazioni a cui partecipano.

### Reti di regolazione genetica

Le piccole molecole necessarie agli organismi biologici, come zuccheri e grassi, sono prodotte nella cellula dalle reazioni chimiche del metabolismo. Le proteine, tuttavia, che sono molecole molto più grandi, sono fabbricate in un modo diverso, seguendo ricette registrate nel materiale genetico della cellula, il DNA.

Le proteine ​​sono molecole a catena lunga formate dalla concatenazione di amminoacidi. La sequenza di amminoacidi proteici è determinata da una sequenza corrispondente memorizzata nel **DNA** della cellula in cui la proteina viene sintetizzata. Questa è la funzione primaria del DNA, che agisce come un mezzo di memorizzazione delle informazioni contenente le sequenze di proteine ​​necessarie alla cellula. Il DNA è esso stesso costituito da unità chiamate **nucleotidi** , di cui esistono quattro specie distinte, indicate con A, C, G e T. Gli aminoacidi in proteine ​​sono codificati nel DNA come trii di nucleotidi consecutivi chiamati **codoni**, come ACG e TTT, e una successione di codoni scandisce la sequenza completa di aminoacidi in una proteina. Un singolo filamento di DNA può codificare per molte proteine ​​e due codoni speciali, chiamati codoni di inizio e di fine, segnalano l'inizio e la fine della codifica di sequenza per una proteina. Il codice del DNA per una singola proteina, dall'inizio alla fine del codone, è chiamato un **gene** .

Le proteine ​​sono create nella cellula da un meccanismo che opera in due fasi. Nel primo stadio, noto come **trascrizione** , un enzima fa una copia della sequenza codificante di un singolo gene. La copia è fatta di RNA, un'altra informazione simile al DNA. Nel secondo stadio, chiamato **traduzione** , la proteina viene assemblata passo dopo passo dalla sequenza RNA. Nel gergo della biologia molecolare, si dice che il gene è stato **espresso** .

In generale, la cellula non ha bisogno di produrre in ogni momento ogni possibile proteina per la quale contiene un gene. Le singole proteine ​​servono a scopi specifici, come la catalizzazione delle reazioni metaboliche, ed è importante che la cellula risponda al suo ambiente attivando o disattivando la produzione di singole proteine. Lo fa con l'uso di **fattori** di **trascrizione** , che sono essi stessi proteine. La presenza nella cellula del fattore di trascrizione per il gene attiva o migliora l'espressione di quel gene, o lo inibisce, a seconda del tipo di fattore di trascrizione (che promuove e inibisce).

Ecco la rete. Essendo proteine, i fattori di trascrizione sono essi stessi prodotti dalla trascrizione dai geni, e il processo di trascrizione è regolato da altri fattori di trascrizione, che di nuovo sono proteine, e così via. L'insieme completo di tali interazioni forma una **rete di regolazione genetica** . I vertici in questa rete sono proteine, o equivalentemente i geni che codificano per loro e un vantaggio diretto dal gene A al gene B indica che A regola l'espressione di B. Si può distinguere tra promuovere e inibire i fattori di trascrizione, dando alla rete due distinti tipi di bordi.

## Reti neurali

Una delle funzioni principali del cervello è quella di elaborare le informazioni e l'elemento principale di elaborazione delle informazioni è il **neurone** , una cellula cerebrale specializzata che combina diversi input per generare una singola uscita.

Un tipico neurone consiste in un corpo cellulare o **soma** , insieme a numerosi tentacoli sporgenti, chiamati **dendriti** , che sono cavi di ingresso per trasportare segnali nella cellula. La maggior parte dei neuroni ha un solo output, chiamato **axon** , che è tipicamente più lungo dei dendriti. Di solito si dirama verso la sua estremità nei **terminali degli assoni** per consentire all'uscita della cella di alimentare l'input di molti altri. C'è un piccolo spazio, chiamato **sinapsi** , tra terminale e dendrite attraverso il quale il segnale di uscita del primo neurone (presinaptico) deve essere trasportato per raggiungere il secondo neurone (postsinaptico).

Al livello più semplice, una rete neurale può essere rappresentata come un insieme di vertici, i neuroni, collegati da due tipi di bordi diretti, uno per gli ingressi eccitatori e uno per gli ingressi inibitori. In pratica, i neuroni non sono tutti uguali. Questa variazione può essere codificata nella nostra rappresentazione di rete da diversi tipi di vertici.

La scienza attuale non può dirci esattamente come il cervello esegue i compiti cognitivi più sofisticati che consentono agli animali di sopravvivere, ma è noto che il cervello cambia costantemente il modello di cablaggio tra i neuroni in risposta agli input e all'esperienza, e si presume che questo modello - la **rete neurale** - contiene gran parte del segreto.

## Reti ecologiche

**Le reti ecologiche** sono reti di interazioni ecologiche tra specie. Le specie in un ecosistema possono interagire in modi diversi: possono mangiare l'un l'altro, possono parassitarsi l'un l'altro, oppure possono avere una qualsiasi varietà di interazioni reciprocamente vantaggiose, come l'impollinazione o la dispersione dei semi.

Le reti alimentari sono le reti ecologiche più studiate. Gli organismi biologici sul nostro pianeta possono essere suddivisi in ecosistemi, gruppi di organismi che interagiscono tra loro e con elementi del loro ambiente. Una **rete alimentare** è una rete diretta che rappresenta le specie preda delle altre in un determinato ecosistema. I vertici della rete corrispondono alle specie e ai bordi diretti alle interazioni predatore-preda. In effetti, gli ecologi disegnano convenzionalmente i bordi nella direzione opposta, dalla preda al predatore, cioè nella direzione del flusso di energia. In alcuni casi sono stati fatti tentativi per misurare non solo la presenza di interazioni tra le specie ma anche la loro **forza**per esempio come la frazione di energia che una specie deriva da ciascuna delle sue prede. Il risultato è una rete orientata ponderata che getta più luce sul flusso di energia attraverso un ecosistema.

Le reti alimentari sono approssimativamente **dirette grafi aciclici** (DAG). La natura aciclica delle reti alimentari indica che esiste una gerarchia intrinseca tra le specie negli ecosistemi. Quelli più in alto nella gerarchia predano quelli più in basso, ma non viceversa, anche se ci sono alcuni controesempi. La posizione di una specie in questa gerarchia è chiamata dagli ecologisti come il suo **livello trofico** . Questo è il **rango** del vertice della specie sul grafico dell'alimento aciclico, ovvero la lunghezza del percorso più lungo che conduce al vertice che rappresenta la specie. Quelle specie in livelli trofici inferiori tendono ad essere più piccole e più abbondanti, mentre quelle in posizioni trofiche più alte sono di solito più grossi e predatori meno numerosi.

Altre reti ecologiche includono reti di ospiti parassiti e reti mutualistiche. **Le reti di parassiti ospiti** sono reti di relazioni parassitarie tra organismi. I parassiti tendono ad avere un corpo più piccolo, quindi i loro ospiti e parassiti possono vivere fuori dai loro ospiti senza ucciderli. Le reti di ospiti-parassiti sono dirette a reti acicliche. **Le reti mutualistiche** sono reti di interazioni reciprocamente vantaggiose tra le specie. Questi includono reti di piante e animali (insetti) che li impollinano, reti di piante e animali (uccelli) che disperdono i loro semi e reti di specie di formiche e le piante che proteggono e mangiano. Le reti mutualistiche possono essere rappresentate come grafici bipartiti non orientati.

# Information Networks

Le reti di informazioni consistono in elementi di dati collegati in qualche modo. Forse l'esempio più noto è il World Wide Web.

## Il World Wide Web

Il Web è una rete in cui i vertici sono pagine Web e i bordi sono i collegamenti ipertestuali che ci consentono di navigare da una pagina all'altra. Più precisamente, esiste un margine dalla pagina P alla pagina Q se la pagina P contiene almeno un collegamento ipertestuale che punta alla pagina Q. In genere, il numero effettivo di collegamenti ipertestuali tra due pagine non è importante e quindi la rete che modella il Web **non è un grafo diretto**. Inoltre, i collegamenti automatici delle pagine vengono ignorati.

Il Web è stato inventato nel 1989 dai fisici **Tim Berners-Lee** presso l'Organizzazione europea per la ricerca nucleare (CERN). Berners-Lee ha proposto l' **HyperText Markup Language** (HTML) per tenere traccia dei dati sperimentali al CERN. Un estratto interessante della proposta originale e lungimirante in cui Berners-Lee tenta di persuadere il management del CERN ad adottare il nuovo sistema globale di ipertesto segue:

Dovremmo lavorare verso un sistema di informazioni universali, in cui la generalità e la portabilità siano più importanti delle tecniche grafiche fantasiose e delle complesse funzioni extra. L'obiettivo sarebbe quello di consentire di trovare un luogo per qualsiasi informazione o riferimento che si riteneva fosse importante e un modo per trovarlo in seguito. Il risultato dovrebbe essere sufficientemente attraente da usare che le informazioni contenute supererebbero una soglia critica.

La struttura del Web era infatti immaginata molto prima. Nel 1945 **Vannevar Bush** scrisse un celebre articolo su The Atlantic Monthlyintitolato "As We May Think" che descrive un dispositivo futuristico chiamato Memex. Bush scrive:

Appariranno nuove forme completamente nuove di enciclopedie, pronte con una trama di scia associativa che le attraversa, pronte per essere lasciate cadere nel Memex e lì amplificate.

La struttura del Web può essere misurata utilizzando un crawler, un programma per computer che naviga automaticamente sul Web alla ricerca di pagine utilizzando una tecnica di ricerca in ampiezza. I web crawler sono tipicamente sfruttati dai **motori di ricerca Web** , come Google, dal momento che conoscere la struttura del Web è fondamentale per fornire agli utenti classifiche significative delle pagine Web. Più in particolare, i moduli coinvolti nel processo di ricerca Web sono brevemente descritti nel seguito:

1. **Crawler** . La natura hyperlinked del Web viene sfruttata dal software di scansione, che genera robot virtuali, detti spider, che saltano ripetutamente da una pagina all'altra seguendo i collegamenti ipertestuali. Le pagine raccolte dagli spider sono temporaneamente archiviate come pagine complete in un repository centrale.
2. **Indicizzatore** . Questo modulo prende la versione non compressa delle pagine dal repository della pagina, estrae da esse solo le informazioni interessanti e le memorizza in varie strutture di dati chiamate indici. Un motore di ricerca Web funziona principalmente con due indici: un **indice di contenuto** , contenente informazioni sul contenuto delle pagine e un **indice di struttura** , che memorizza preziose informazioni sulla topologia del collegamento ipertestuale del Web. L'indice del contenuto utilizza una struttura di file invertita. Questo file è simile a un indice di un libro e contiene, accanto a ciascun termine significativo, un elenco di posizioni di pagina in cui compare il termine. Inoltre, l'indice contiene il numero di occorrenze del termine in ogni pagina e se il termine appare o meno in sezioni di pagina importanti come il titolo.
3. **Processore di query** . I moduli sopra descritti esistono e operano indipendentemente dalle query dell'utente. Il processore di query è, invece, un modulo dipendente dalla query che accetta le query degli utenti e risponde in tempo reale. Il modulo di query analizza la query e seleziona dall'indice di contenuto le pagine rilevanti per la query posizionata.
4. **Modulo di classificazione** . Questo pezzo finale di un motore di ricerca Web ha il compito cruciale di classificare le pagine rilevanti della query mettendo le prime pagine che l'utente desidera maggiormente. In genere, ciascuna pagina pertinente viene valutata utilizzando un **punteggio del contenuto** , che misura quanto è simile una pagina rispetto alla query dell'utente, ad esempio contando il numero di occorrenze dei termini della query all'interno della pagina e controllando se i termini della query si verificano in sezioni di pagina importanti come il titolo. Inoltre, per ogni pagina viene calcolato un **punteggio di importanza** , che misura lo stato della pagina all'interno dell'intera rete Web. Il punteggio di importanza calcolato dal motore di ricerca di Google è ben noto come **PageRank**punteggio della pagina; l'idea è di dare più importanza alle pagine che sono collegate da altre pagine importanti. I punteggi di contenuto e importanza vengono quindi combinati in un punteggio complessivo e le pagine vengono presentate all'utente in base a questo valore globale.

## Reti di citazioni accademiche

La maggior parte dei documenti accademici fa riferimento ad altri documenti precedenti, di solito in una bibliografia alla fine del documento. Le reti di citazioni accademiche sono reti in cui i nodi sono unità bibliometriche, come giornali e riviste, ei bordi diretti sono le citazioni bibliografiche tra di loro.

Due esempi tipici sono le **reti di citazioni di articoli** , in cui i vertici sono carte e i bordi sono le citazioni bibliografiche tra articoli e **reti di citazioni su riviste**, in cui i nodi sono riviste accademiche e i bordi sono le citazioni bibliografiche tra articoli pubblicati nelle riviste. Nelle reti di citazioni di articoli, il numero effettivo di citazioni tra due articoli viene ignorato e la rete viene rappresentata come un grafico diretto non ponderato. Inoltre, dal momento che le citazioni su carta puntano indietro nel tempo, dai nuovi ai vecchi documenti, le reti di citazioni di articoli sono grafi aciclici. D'altra parte, le reti di citazioni di riviste sono grafici diretti ponderati, in cui i bordi sono ponderati con il numero effettivo di citazioni tra i giornali coinvolti. Poiché le riviste contengono articoli pubblicati in periodi diversi, le reti di citazioni di riviste contengono in genere loop.

Le citazioni sono spesso un modo di mostrare il debito intellettuale perché le idee e i risultati del documento precedente sono stati utili per le indagini fornite nel vecchio documento. Quindi, il numero di citazioni ricevute da un documento di altri giornali fornisce un'indicazione **dell'impatto** di un documento all'interno della comunità accademica. L'impatto delle carte scritte da uno studioso può essere usato come un criterio per la valutazione della qualità della ricerca effettuata dallo studioso.

Recentemente, sono stati proposti indicatori bibliometrici, come l' [Eigenfactor](http://www.eigenfactor.org/) , che sfruttano la topologia della rete di citazioni del journal per calcolare i punteggi di importanza per le riviste. L'indicatore Eigenfactor funziona in modo simile al PageRank per le pagine Web, dando più importanza alle riviste citate da altre riviste importanti. Quindi, le riviste accademiche all'interno di un determinato campo possono essere classificate in base alla loro importanza e queste classifiche possono essere sfruttate, ad esempio, dai bibliotecari per acquistare le riviste più importanti all'interno di un budget limitato.

Le due principali fonti di dati bibliometrici commerciali sono: **Web of Science** , di proprietà di Thomson-Reuters e **Scopus** , di Elsevier. Sia Web of Science che Scopus sono database proprietari basati su sottoscrizioni gestiti da personale professionale. D'altra parte, [Google Scholar](http://scholar.google.com/)utilizza l'indicizzazione automatica delle citazioni e fornisce l'accesso gratuito.

Le reti di citazioni finora descritte non sono l'unica rappresentazione in rete dei modelli di citazione. Una rappresentazione alternativa è la **rete di cocitazione** . Si dice che due documenti siano cocitati se entrambi sono citati dallo stesso terzo foglio. Il Cocitation è spesso considerato un indicatore del fatto che i documenti trattano argomenti correlati. Una rete di cocitazione è una rete in cui i vertici rappresentano le carte e i bordi rappresentano la cocitazione di coppie di carte. Si noti che la cocitazione è una relazione simmetrica, quindi i bordi delle reti di cocitazione non sono diretti. È anche possibile fornire una forza per la cocitazione tra due documenti come il numero di altri documenti che citano entrambi e quindi creare una rete di cocitazione ponderata.

Un altro concetto correlato è l' **accoppiamento bibliografico** . Si dice che due documenti siano accoppiati bibliograficamente se citano la stessa altra carta. L'accoppiamento bibliografico, come la cocitazione, può essere preso come un indicatore che i documenti trattano di materiale correlato. Si può definire una rete di accoppiamento bibliografico, ponderata o meno, in cui i vertici sono carte i bordi non orientati indicano l'accoppiamento bibliografico.

Le reti di citazioni, in particolare le reti di citazioni di articoli, sono per molti versi simili al World Wide Web. I vertici della rete contengono informazioni sotto forma di testo e immagini, proprio come fanno le pagine Web, e i collegamenti da un foglio a un altro svolgono un ruolo simile ai collegamenti ipertestuali sulle pagine Web, allertando il lettore quando le informazioni relative all'argomento di un foglio può essere trovato su un altro. Le carte con molte citazioni sono spesso più influenti di quelle con pochi, proprio come nel caso delle pagine Web che sono collegate a molte altre pagine; in particolare, l'indicatore bibliometrico Eigenfactor è stato progettato per funzionare nella rete di citazioni del journal come il PageRank fa sulla rete Web. Finalmente,

Tuttavia, gli studi quantitativi sulle reti di citazioni sono molto più vecchi del Web e risalgono agli anni '60 con gli studi di **de Solla Price** . In particolare, Sergey Brin e Larry Page sono stati ispirati dai lavori sulle reti di citazioni nello sviluppo dell'algoritmo PageRank per Google. Gli studi su pubblicazioni e citazioni accademiche rientrano nel campo della **bibliometria** , una branca delle scienze dell'informazione e della biblioteca.

## Reti di raccomandazione

Un tipo di rete di informazioni importante per la tecnologia e il commercio è la rete di raccomandazione. **Le reti di raccomandazione**rappresentano le preferenze delle persone per le cose, ad esempio per i prodotti venduti da un rivenditore. I commercianti online, ad esempio, tengono un registro dei clienti che hanno acquistato i prodotti e, a volte, chiedono loro se apprezzano i prodotti. Come altro esempio, molte grandi catene di supermercati registrano gli acquisti effettuati da ciascuno dei loro clienti abituali, di solito identificati da una carta di codici a barre che viene scansionata durante gli acquisti.

La rappresentazione appropriata di una rete di raccomandazione è come un **grafico bipartito** in cui i due tipi di vertici rappresentano prodotti e clienti, con spigoli che collegano i clienti ai prodotti che acquistano o che preferiscono. Si possono aggiungere anche pesi ai bordi per indicare quanto spesso la persona ha acquistato il prodotto di come la persona piace.

Il principale interesse commerciale nelle reti di raccomandazione deriva dal loro utilizzo nei **sistemi di raccomandazione** . Si tratta di algoritmi informatici che cercano di indovinare gli elementi che le persone apprezzeranno confrontando le preferenze conosciute di una persona con quelle di altre persone che hanno acquistato prodotti simili. Se la persona A ha acquistato o apprezza molti degli stessi prodotti delle persone B, C e D, e se le persone B, C e D sono state acquistate o hanno acquistato un altro oggetto che A non ha mai acquistato o espresso un'opinione, allora forse A piacerebbe anche quell'oggetto Il sito Web Amazon.com , ad esempio, utilizza i sistemi di raccomandazione per raccomandare i titoli dei libri e molti supermercati stampano buoni sconto per i clienti per prodotti che il cliente non ha mai acquistato ma potrebbe essere interessato a provare.

# Social Networks

I social network sono reti in cui i nodi sono persone, o talvolta gruppi di persone, e i bordi rappresentano una qualche forma di interazione sociale tra loro, come l'amicizia. I sociologi hanno sviluppato il proprio linguaggio per discutere di reti: si riferiscono ai vertici, alle persone, come **attori** e ai bordi come **legami** .

Per la maggior parte delle persone le parole social network significano servizi di social networking come Facebook e Twitter . In effetti, lo studio dei social network risale molto più lontano. Tra i ricercatori che studiano le reti, i **sociologi** hanno la tradizione più lunga e consolidata di studio empirico quantitativo.

Il vero fondamento del campo è attribuito allo psichiatra **Jacob Moreno** , un immigrato rumeno in America che negli anni '30 si interessò alla dinamica delle interazioni sociali di gruppi di persone. Moreno ha pubblicato nel 1934 un libro intitolato **Who Shall Survive?**che conteneva i semi del campo della **sociometria** , che in seguito divenne l'analisi dei social network. Moreno chiamava i suoi diagrammi **sociogrammi** , piuttosto che i social network. Il primo esempio di social network era un'immagine disegnata a mano che mostrava modelli di amicizia tra i ragazzi e le ragazze in una classe di scolari. La figura rivela che ci sono molte amicizie tra due ragazzi e due ragazze, ma poche tra un ragazzo e una ragazza.

Un altro studio iniziale sui social network è la rete di affiliazione del cosiddetto **Southern Women Study** , pubblicato nel 1941 in un libro intitolato Deep South . Davis, Gardner e Gardner hanno usato i resoconti dei giornali sull'apparizione pubblica delle donne della società per studiare un social network di 18 donne in una città del sud americano. Hanno preso un campione di 14 eventi sociali frequentati dalle donne in questione e registrato quali donne hanno partecipato a quali eventi. Le donne in questa rete possono essere considerate collegate se hanno partecipato allo stesso evento. Una rappresentazione alternativa e più completa dei dati è una rete di affiliazione o un **grafico bipartito**, una rete con due tipi di vertice, che rappresentano le donne e gli eventi, con spigoli che collegano ogni donna agli eventi a cui ha partecipato. Le donne sono state trovate dai ricercatori per dividersi in due sottogruppi, gruppi di conoscenti strettamente uniti con solo interazioni tra i cluster piuttosto lenti.

Da Moreno e Davis et al., L'analisi dei social network è stata applicata a una varietà di comunità diverse, inclusi modelli di amicizia e conoscenza nella popolazione, tra studenti o scolari, contatti tra uomini d'affari e altri professionisti, consigli di amministrazione di società, collaborazioni di scienziati, attori cinematografici e musicisti, reti di contatti sessuali e modelli di incontri, reti criminali come reti di tossicodipendenti e terroristi, reti storiche, social network online e reti sociali di animali.

## Misurare un social network

Una questione cruciale è lo studio dei social network, il metodo empirico per accumulare dati sulla rete. Le due tecniche sono le più utilizzate: **interrogazione diretta** dei soggetti e uso di **documenti d'archivio** .

Il metodo più comune è semplicemente quello di porre domande alle persone. Se sei interessato alle reti di amicizia, ad esempio, chiedi alle persone chi sono i loro amici. La domanda può assumere la forma di interviste dirette con i partecipanti o il completamento da parte dei partecipanti di questionari. I principali svantaggi degli studi sulle reti basati sull'interrogazione diretta dei partecipanti sono che sono prima laboriosi e secondariamente imprecisi. La somministrazione di interviste o questionari e la raccolta di risposte è un lavoro impegnativo. Per questo motivo, la maggior parte degli studi è stata limitata a poche decine o al massimo a centinaia di attori. Inoltre, le risposte fornite dagli intervistati sono sempre, in una certa misura, soggettive. Se chiedi alle persone chi sono i loro amici, persone diverse interpretano l'amicizia in modi diversi e quindi danno diversi tipi di risposte.

Un aumento importante, voluminoso, e aver visto spesso altamente affidabile fonte di dati rete sociale è **documenti d'archivio** . Tali registrazioni sono spesso impressionanti nella loro scala che ci consente di costruire reti di grandi dimensioni. Un caso speciale importante della ricostruzione di reti da record di archivio è la **rete di affiliazione**. Una rete di affiliazione è una rete in cui gli attori sono collegati tramite la co-iscrizione di gruppi di qualche tipo. La rappresentazione più completa di una rete di affiliazione è come un grafico bipartito, in cui le reti hanno due tipi di vertice che rappresentano gli attori e i gruppi, con spigoli che collegano gli attori ai gruppi a cui appartengono. Forse l'esempio più noto è la rete di collaborazione di attori cinematografici, in cui gli attori nel senso della rete sono gli attori in senso drammatico, ei gruppi a cui appartengono sono i cast del film. La rete è la base di un popolare gioco di società, a volte chiamato **Six Degrees di Kevin Bacon** , in cui si cerca di collegare un attore a Kevin Bacon attraverso catene di costar intermedie. Un altro esempio di una grande rete di affiliazione è il**rete** di **collaborazione** degli accademici. In questa rete un attore è un autore accademico e un gruppo è l'insieme di autori di una carta appreso. Archivi eccellenti e molto completi esistono in molti campi accademici, dai quali si possono assemblare e studiare grandi reti di collaborazione.

Un altro modo per misurare una rete è **l'osservazione diretta** : semplicemente osservando le interazioni tra gli attori è possibile, per un periodo di tempo, formare un'immagine delle reti di legami invisibili che esistono tra di loro. Un'arena in cui l'osservazione diretta è essenzialmente l'unica tecnica sperimentale possibile per assemblare una rete è lo studio delle **reti sociali degli animali** - chiaramente gli animali non possono essere esaminati tramite interviste o questionari. Sono stati condotti studi informativi per scimmie, canguri e delfini.

Una rete può cambiare nel tempo e talvolta i dati di rete vengono **risolti nel tempo** : la data di ogni interazione tra coppie di vertici, che forma un bordo della rete, viene registrata. Ad esempio, i dati della rete di collaborazione sono spesso risolti nel tempo poiché le bibliografie contengono almeno l'anno di ciascuna pubblicazione accademica registrata. Quindi, i collegamenti di collaborazione tra gli autori possono essere timbrati con l'anno in cui la collaborazione ha avuto luogo. Gli studi di rete risolti in tempo, o studi **longitudinali** , come vengono chiamati in sociologia, consentono un'analisi temporale della rete, ovvero le proprietà della rete modificate nel tempo.

## Campionamento dei social network

Esistono due tecniche principali per campionare una popolazione: **campionamento a palle di neve** , che richiama una **ricerca** di un grafico al **primo respiro** e **un campionamento casuale** , simile a una **ricerca di profondità** su un grafico.

Nel campionamento a palle di neve, gli investigatori sondano la popolazione facendo in modo che alcuni membri forniscano i dettagli di contatto per gli altri. Puoi trovare un membro iniziale della popolazione di interesse e intervistarli. Quindi, una volta raggiunta la fiducia, invitali anche a nominare altri membri della popolazione target con cui sono a conoscenza. Quindi intervisti tutti quei conoscenti che chiedono di nominare ulteriori contatti, e così via attraverso una serie di ondate di campionamento.

Un'alternativa al campionamento della palla di neve è il campionamento **casuale** , simile a una **ricerca in profondità**su un grafico. In questo metodo si inizia nuovamente con un singolo membro della comunità di destinazione, intervistandoli e determinando i loro contatti. Quindi, invece di intervistare tutti questi contatti, uno ne sceglie uno a caso e ne intervista solo uno ad ogni passaggio. Se la persona in questione non può essere trovata o rifiuta di essere intervistata, si sceglie semplicemente un altro contatto. È tuttavia molto importante che si determinino realmente tutti i contatti di ciascun individuo, anche se la maggior parte delle volte viene perseguito solo uno di essi. Questo perché il metodo per funzionare correttamente deve fare una scelta casuale tra i contatti. La sequenza risultante di contatti corrisponde a una passeggiata casuale sul social network di interesse.

Entrambi i metodi di campionamento introducono **pregiudizi** nella scelta del campione. Tuttavia, nel campionamento random-walk possiamo facilmente correggere questo bias di campionamento. In effetti, la probabilità di campionamento asintotico dei vertici in una passeggiata casuale è semplicemente proporzionale al **grado dei vertici** . Vertici di alto livello, cioè con molti vicini, hanno più probabilità di essere visitati da una passeggiata casuale perché ci sono più modi per raggiungerli. D'altra parte, nel limite di un gran numero di onde di campionamento, campioni di palle di neve campionano attori con probabilità proporzionali alla loro **centralità dell'autovettore**. Vertici di alto livello, ma anche quelli con pochi vicini di alto livello, hanno più probabilità di essere visitati durante un processo di campionamento della palla di neve. Mentre possiamo determinare il grado dei vertici come parte del processo di campionamento, il calcolo della centralità dell'autovettore richiede una conoscenza completa della rete, che per definizione non abbiamo. Inoltre, nel campionamento della palla di neve, la dimensione del campione cresce esponenzialmente con il numero di onde di campionamento e quindi in genere si esegue solo un numero logaritmico di onde, il che non è sufficiente per il processo di campionamento per raggiungere l'equilibrio. Nel campionamento casuale la dimensione del campione cresce linearmente, e quindi il regime asintotico viene raggiunto abbastanza rapidamente per campioni di dimensioni relativamente piccole.

# Technological Networks

<http://users.dimi.uniud.it/~massimo.franceschet/teaching/datascience/network/technological.html>

Le reti tecnologiche sono reti **fisiche** che costituiscono la spina dorsale delle moderne società tecnologiche. Questi includono Internet, reti telefoniche, reti elettriche, reti di trasporto e reti di distribuzione.

## La rete

Internet è la rete mondiale di connessioni fisiche tra computer e dispositivi correlati. Si tratta di una rete di dati a **commutazione di pacchetto** , il che significa che i messaggi inviati su di esso vengono suddivisi in pacchetti, piccoli blocchi di dati, che vengono inviati separatamente sulla rete e ricomposti nuovamente in un messaggio completo all'altra estremità.

Il formato dei pacchetti segue uno standard noto come IP ( **Internet Protocol** ) e include un indirizzo IP in ogni pacchetto che specifica la destinazione del pacchetto, in modo che possa essere instradato correttamente attraverso la rete.

L'utilizzo di un modello di commutazione a pacchetto per Internet consente ai computer di trasmettere e ricevere dati a intermittenza oa velocità variabile senza accumulare capacità sulla rete. Inoltre, consente una certa quantità di inaffidabilità nella rete. Non è raro che i pacchetti spariscano su Internet e non raggiungano mai la loro destinazione, spesso perché i pacchetti vengono deliberatamente eliminati per ridurre la congestione nelle parti più trafficate della rete. Solo i pacchetti del messaggio che sono stati persi devono essere nuovamente inviati per completare il messaggio. Un protocollo software denominato **Transport Control Protocol** (TCP), con le esecuzioni su IP, esegue automaticamente la verifica degli errori e la ritrasmissione necessari, senza la necessità di intervento da parte di utenti di computer o altri software.

La rappresentazione di rete più semplice di Internet è quella in cui i vertici della rete rappresentano computer o altri dispositivi e i bordi rappresentano connessioni fisiche tra di loro, come linee in fibra ottica o connessioni wireless. In realtà, i computer ordinari occupano solo la periferia della rete; sono le destinazioni o le fonti del flusso di dati, ma non fungono da punti intermedi per il flusso di dati tra gli altri. I nodi interni di Internet sono i **router** , potenti macchine speciali per gli incroci tra le linee dati che ricevono i pacchetti di dati e li inoltrano verso la destinazione prevista.

La forma generale generale di Internet è composta da tre cerchi di vertici. Il cerchio più interno, il cuore della rete, è costituito da router ad alte prestazioni e linee a lunga banda larga. Queste linee principali sono le autostrade di Internet e questo cerchio è chiamato la spina dorsale di Internet. Il backbone è gestito da **fornitori di backbone di rete** (NBP). La seconda cerchia di Internet è composta da **provider di servizi Internet** (ISP), che stipulano contratti con NBP per fornire la connessione agli **utenti finali** , i consumatori finali della larghezza di banda Internet, che formano la terza cerchia di Internet.

La struttura di rete di Internet non è dettata da alcuna **autorità centrale** . In particolare, non è necessario rivolgersi a qualsiasi autorità Internet centrale per ottenere il permesso di costruire un nuovo stimolo su Internet o per renderne uno fuori servizio. Lo schema per il routing dei pacchetti su Internet è progettato in modo tale che se vengono aggiunti nuovi bordi o vertici o vengono rimossi quelli vecchi, i router prenderanno nota e adegueranno la loro politica di routing in modo appropriato.

Per fare una rappresentazione affidabile di Internet, dobbiamo misurare la struttura di Internet in qualche modo. Ci sono due metodi principali per farlo: traceroute e tabelle di routing. **Traceroute** è uno strumento standard per scoprire il percorso seguito dai pacchetti di dati che viaggiano tra il nostro computer (o qualsiasi computer a cui abbiamo accesso) e molti altri su Internet. Oltre a un indirizzo di destinazione, ogni pacchetto contiene un indirizzo di origine e un **time-to-live** (TTL). Il TTL è un numero che specifica il numero massimo di passaggi che un pacchetto può eseguire per raggiungere la sua destinazione. Se il TTL di un pacchetto raggiunge lo zero, il pacchetto viene scartato e un messaggio viene inviato al computer di origine informando che il pacchetto è stato scartato e dove è arrivato.

Possiamo utilizzare questo meccanismo per tracciare il percorso seguito da un dato imballato dal nostro computer a uno di destinazione. Inviamo ripetutamente un pacchetto con l'indirizzo di destinazione che ci interessa e un TTL crescente (1, 2, 3 e così via). Supponendo che ogni pacchetto prenda la stessa rotta verso la destinazione, questo potrebbe non essere vero se la congestione si verifica e la rete reindirizza il pacchetto con connessioni meno congestionate, quindi la procedura genera una sequenza di indirizzi IP che specifica il percorso dalla sorgente alla destinazione. L'idea è di assemblare una grande serie di dati di percorsi traceroute tra molte coppie di vertici differenti. Viene selezionato per la prima volta un numero di computer sorgente, idealmente ben distribuiti sulla rete. Per ogni origine, il percorso verso un numero di computer di destinazione viene calcolato utilizzando traceroute, ottenendo una struttura ad albero radicata al vertice sorgente. Quindi, una semplice unione di queste strutture ad albero fornisce un'istantanea (non necessariamente completa) della rete.

La maggior parte degli studi su Internet ignorano i computer degli utenti finali e si limitano ai soli router, concentrandosi sulla zona interna della rete (backbone e ISP). Tali mappe di Internet sono rappresentazioni a **livello di router** . Altre rappresentazioni a grana grossa vengono create raggruppando insiemi di indirizzi IP in singoli vertici. Questi includono rappresentazioni a livello di sottoreti, domini e sistemi autonomi.

Una **sottorete** è un insieme di indirizzi IP definiti come segue. Gli indirizzi IP sono composti da quattro numeri, ognuno compreso tra 0 e 255, come 141.211.144.190. Gli indirizzi del modulo 141.211.144.xxx formano una subnet di classe C, gli indirizzi del modulo 141.211.xxx.yyy formano una sottorete di classe B e gli indirizzi del modulo 141.xxx.yyy.zzz formano una sottorete di classe A. Poiché tutti gli indirizzi in una sottorete di classe C sono solitamente assegnati alla stessa organizzazione, un modo ragionevole di dati di rete Internet a grana grossa consiste nel raggruppare i vertici in sottoinsiemi di classe C e posizionare un bordo tra due subnet se qualsiasi router in uno ha una connessione a qualsiasi router nell'altro. Un **dominio** è un gruppo di computer e router sotto il controllo di una singola organizzazione e identificato da un nome di dominio simileuniud.it. Infine, un **sistema autonomo** è simile a un dominio: si tratta di una raccolta di computer e router, solitamente sotto un unico controllo amministrativo, all'interno del quale il routing dei dati viene gestito indipendentemente dalla più ampia Internet, da cui il nome sistema autonomo.

La granulosità grossolana a livello di sistema autonomo non viene eseguita con dati derivati ​​dal campionamento traceroute ma con dati basati su **tabelle di routing**. I router Internet mantengono tabelle di routing che consentono loro di decidere in quale direzione devono essere inviati i pacchetti in entrata per raggiungere al meglio la loro destinazione. Consistono in elenchi di percorsi completi dal router in questione verso destinazioni su Internet. In teoria i router hanno bisogno di memorizzare solo il primo passo del percorso; in pratica salvano l'intero percorso verso la destinazione per il calcolo efficiente dei percorsi. In effetti, le tabelle di routing nei router sono rappresentate a livello di sistemi autonomi. Questo è sufficiente poiché i dati che passano all'interno del sistema autonomo sono gestiti autonomamente dal sistema, mentre i dati che passano tra sistemi autonomi vengono gestiti a livello di Internet utilizzando il Border Gateway Protocol (BGP). Per ottenere un'istantanea ragionevole di Internet, viene raccolto un set di tabelle router, ciascuna contenente un numero elevato di percorsi a partire da un singolo sistema autonomo contenente il router di origine. L'unione di queste tabelle router fornisce un'istantanea non necessariamente completa di Internet a livello di sistema autonomo. Come con traceroute, è importante che i router di origine siano ben distribuiti sulla rete per evitare la duplicazione dei percorsi e del lavoro.

## La rete telefonica

La rete telefonica è composta da linee fisse e collegamenti wireless che trasmettono le chiamate telefoniche. È importante rendersi conto che anche le chiamate effettuate sui telefoni wireless vengono ancora trasmesse principalmente sulla rete telefonica tradizionale. Il segnale proveniente da un telefono senza fili fa il primo passo del suo viaggio in modalità wireless verso una vicina torre di trasmissione, ma da lì percorre le normali linee telefoniche.

Al contrario di Internet, la rete telefonica tradizionale non è affollata. Invece, la rete telefonica è **commutata a circuito** , il che significa che la compagnia telefonica ha un numero di linee o circuiti disponibili per effettuare chiamate telefoniche e assegna quindi a singoli chiamanti. Nei primi giorni dei sistemi telefonici le linee erano in realtà cavi singoli, uno per ogni chiamata che l'azienda poteva trasportare. In seguito, le compagnie telefoniche hanno utilizzato tecniche per il **multiplexing dei** segnali telefonici, ovvero inviando contemporaneamente più chiamate sullo stesso filo. L'eccezione è l'ultimo miglio di connessione al singolo abbonato, che di solito porta una telefonata alla volta.

La rete telefonica ha un **design a tre livelli** . I singoli abbonati telefonici sono collegati tramite linee locali a centrali telefoniche locali, che sono collegate su linee urbane condivise agli uffici a lunga distanza. Gli uffici a lunga distanza sono quindi collegati tra loro da ulteriori linee principali. Questa struttura è, per molti aspetti, piuttosto simile a quella di Internet. Le chiamate telefoniche locali possono essere gestite dalla sola centrale locale e non è necessario utilizzare alcuna linea urbana.

Oggi la maggior parte della rete telefonica non è più commutata. Invece, è un pacchetto digitale a commutazione di rete come Internet, con chiamate vocali digitalizzate, suddivise in pacchetti e trasmesse tramite telefono o collegamenti in fibra ottica. Solo l'ultimo miglio di linee è ancora trasportato su un circuito dedicato vecchio stile, e anche la situazione sull'ultimo miglio sta cambiando con l'avvento dei **servizi di telefonia Internet** (voce su IP), che inviano chiamate vocali su Internet piuttosto che su linee telefoniche dedicate. Quindi Internet e la rete telefonica non sono reti disgiunte e potrebbe non passare molto tempo prima che il telefono e le reti Internet si uniscano in un'unica rete.

## Reti elettriche

Una rete elettrica, in questo contesto, è la rete di linee di trasmissione ad alta tensione che forniscono il trasporto a lunga distanza di energia elettrica all'interno e tra i paesi. Le linee di alimentazione locali a bassa tensione sono normalmente escluse. I vertici di una rete elettrica corrispondono alle stazioni di generazione e alle sottostazioni di commutazione e i bordi corrispondono alle linee di alta tensione.

I guasti nelle reti elettriche possono avere **effetti a cascata** : l'errore di un nodo può provocare ricorsivamente l'errore dei nodi connessi. Questo comportamento può dar luogo a risultati sorprendenti come la distribuzione osservata della legge di potenza nelle dimensioni delle interruzioni di corrente.

## Trasporti e reti di distribuzione

Le reti di trasporto comprendono rotte aeree e reti stradali e ferroviarie. Le reti di distribuzione comprendono oleodotti e gasdotti, linee idriche e fognarie e percorsi utilizzati dall'ufficio postale e dalle società di consegna dei pacchi.

Una classe di reti di distribuzione che è stata relativamente ben studiata sono le reti fluviali, dove i bordi sono fiumi o flussi e i vertici sono le loro intersezioni. Di particolare rilievo è il fatto che le reti fluviali, per un'eccellente approssimazione, assumono la forma di **alberi** . Cioè, non contengono loop, quindi l'acqua in qualsiasi punto della rete scorre via attraverso un singolo percorso. Simili per alcuni aspetti alle reti fluviali sono le reti di vasi sanguigni negli animali e il loro equivalente nelle piante.