FONDAZIONE
EDMUND
MACH

Consiglio Nazionale delle Ricerche

gene@home

TN-Grid BOINC platform

Hello World

UNIVERSITÀ DEGLI STUDI
DI TRENTO

Dipartimento di Ingegneria
e Scienza dell'Informazione

Prof. Enrico Blanzieri

Professore al Dipartimento di Ingegneria e Scienza dell'Informazione dell'Università degli Studi di Trento

Docente del corso di
"Laboratory of Biological Data-Mining"

Hello World

Dott. Claudio Moser

A capo del gruppo di ricerca sulle Funzioni Geniche per la Fondazione Edmund Mach di S. Michele all'Adige

FONDAZIONE EDMUND MACH

Docente nel corso di *"Laboratory of Biological Data-Mining"* per la parte biologica

Hello World

Consiglio Nazionale delle Ricerche

Dott. Valter Cavecchia

Esperto della piattaforma BOINC, nonché *“Evangelist”* su BOINC.Italy

Collaboratore Tecnico degli Enti di Ricerca, Affiliato all'Istituto dei Materiali per l'Elettronica ed il Magnetismo e responsabile dei servizi informatici dell'Area di Ricerca di Trento

Outline

- Chi siamo
- gene@home
- BOINC
- Implementazione
- Futuro

Chi siamo

Francesco Asnicar

I.T.I.S. “V. E. Marzotto”

Laurea triennale in Scienza dell’Informazione

Master student in bio-informatics

Chi siamo

Nadir Sella

Liceo Scientifico N. Tron

Laurea triennale in Scienza dell'Informazione

Master student in bio-informatics

Outline

- Chi siamo
- gene@home
  - Introduzione biologica
  - Obiettivo biologico
  - PC
  - PC-IM
- BOINC
- Implementazione
- Futuro

gene@home

Progetto di

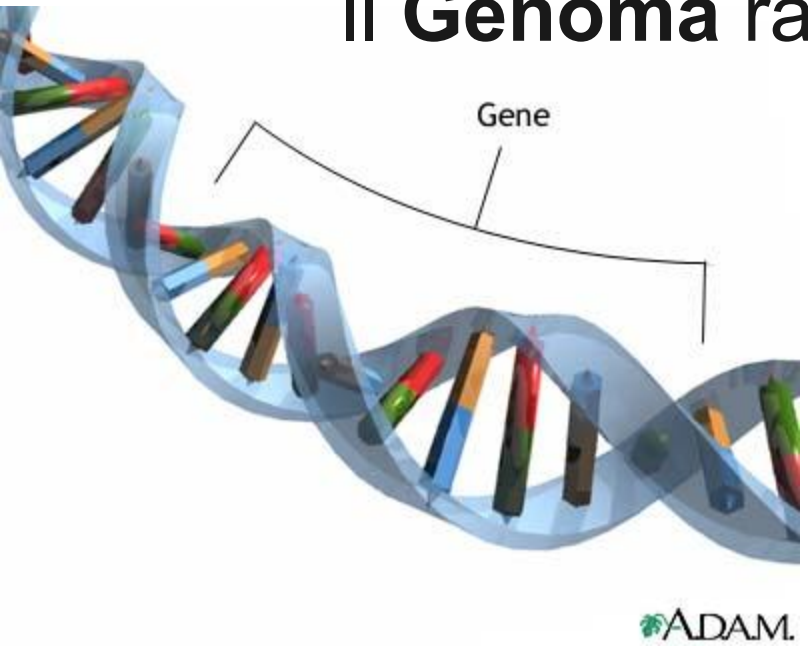
biologia computazionale distribuita

Introduzione biologica

Il **Gene** identifica una porzione di DNA, che corrisponde a una specifica proteina

Il **Genoma** rappresenta l'insieme dei geni di un dato organismo

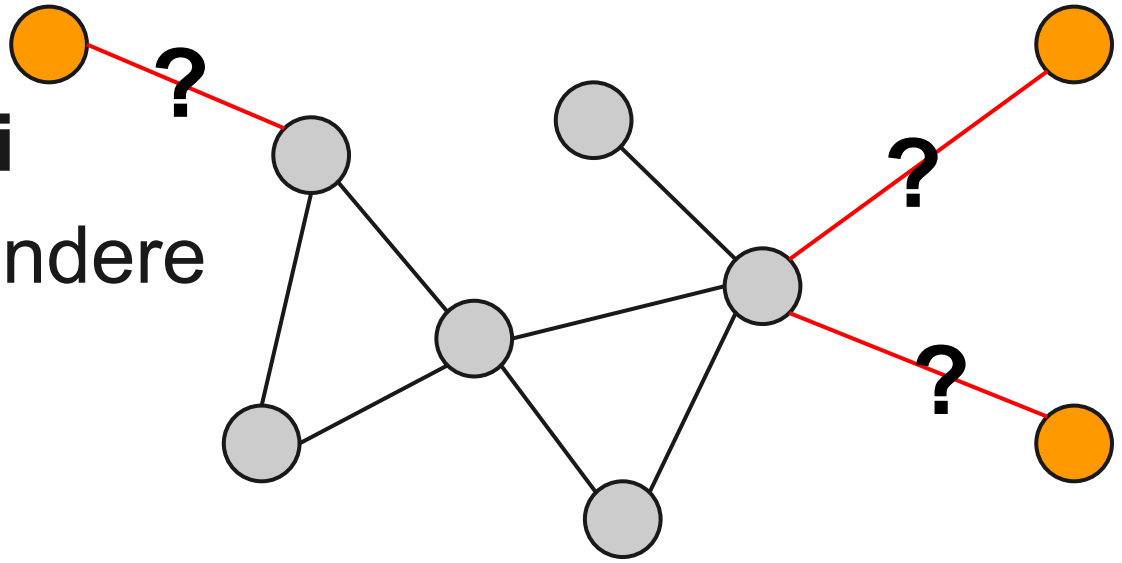
Le interazioni tra un insieme di geni costituisce una **Rete Genica**



Obiettivo biologico

Le **relazioni tra geni** sono poco conosciute!

Vogliamo cercare **relazioni** tra geni per espandere una rete genica conosciuta



Obiettivo biologico

Cosa significa *“trovare relazioni tra geni”*?

Ogni cellula contiene lo **stesso genoma**, allora perchè pelle e unghie sono **diverse**?

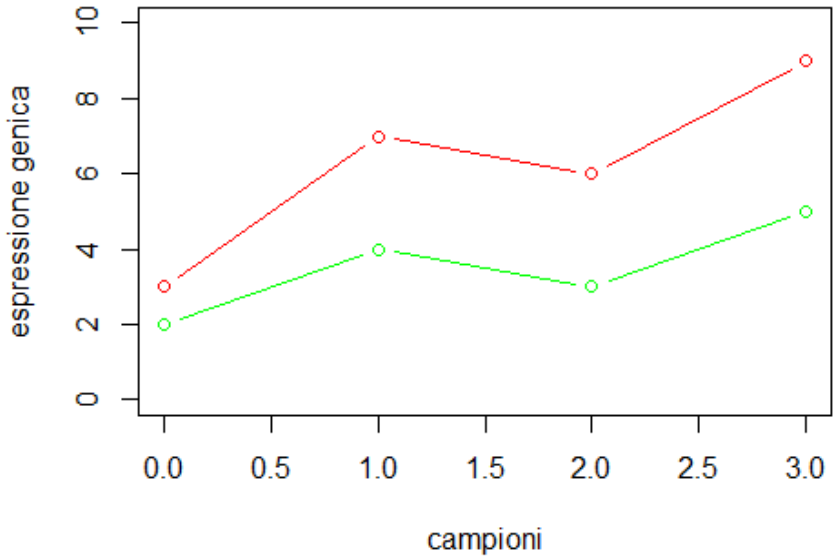
A causa del **livello di Espressione Genica**!

Misura la quantità di proteine che un gene produce in una cellula

Obiettivo biologico

Significa *"paragonare i livelli di espressione genica di due geni"*

La relazione diventa **correlazione** quando le due curve hanno un andamento simile



PC

Le **correlazioni** servono per scoprire e approfondire alcuni importanti processi biologici

A lato pratico l'algoritmo considera i geni tutti correlati tra loro, e toglie via via gli archi tra geni **indipendenti**

PC

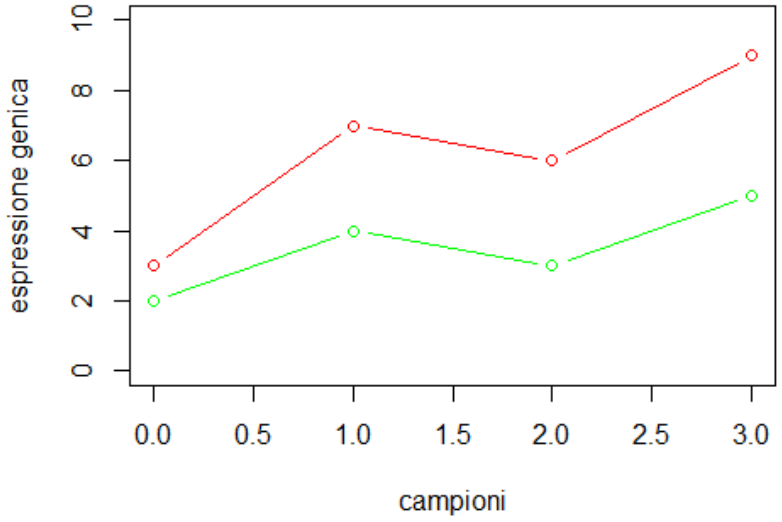
L’algoritmo di **P**eter-**C**lark cerca **indipendenza** tra geni utilizzando i livelli di **espressione genica** di ciascun gene in diversi campioni

Due nodi sono **indipendenti** se poco correlati

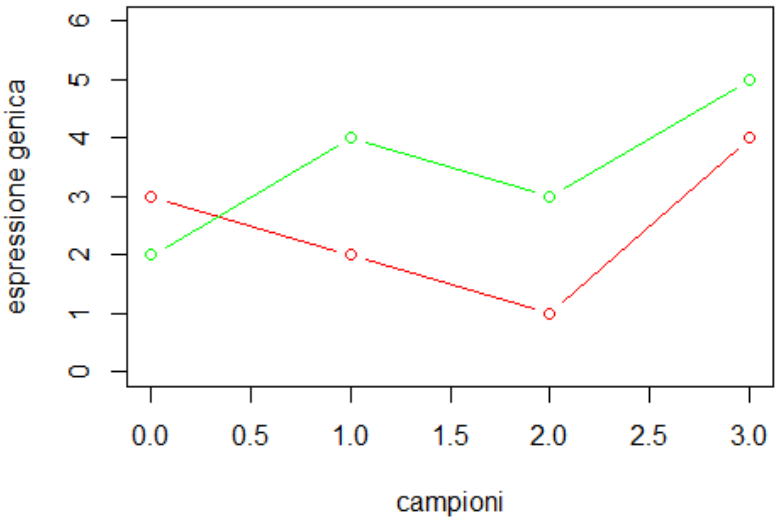
La correlazione tra due geni viene misurata con il **coefficiente di Pearson**

$$r = \frac{\operatorname{cov}(X, Y)}{s_x s_y}$$

PC



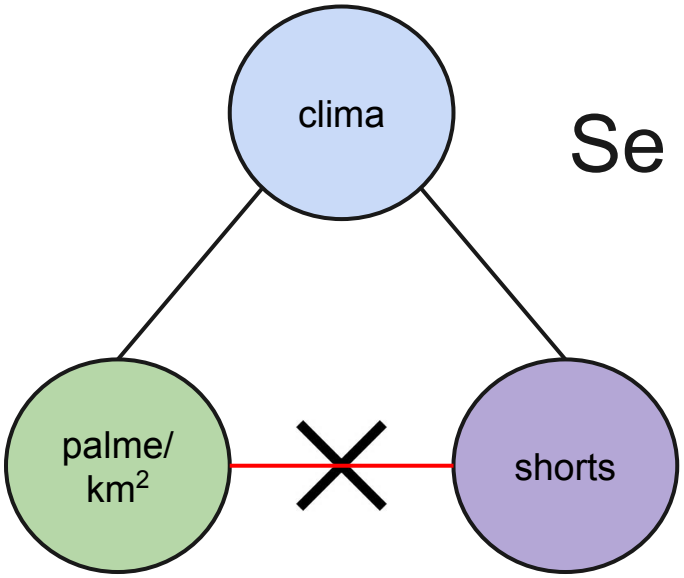
Coefficiente di Pearson = 0.98



Coefficiente di Pearson = 0.40

PC

Due nodi, inizialmente dipendenti, possono diventare indipendenti a causa di altri nodi



Se si aggiunge la conoscenza di *“clima”*, *“palme”* e *“shorts”* diventano indipendenti

PC

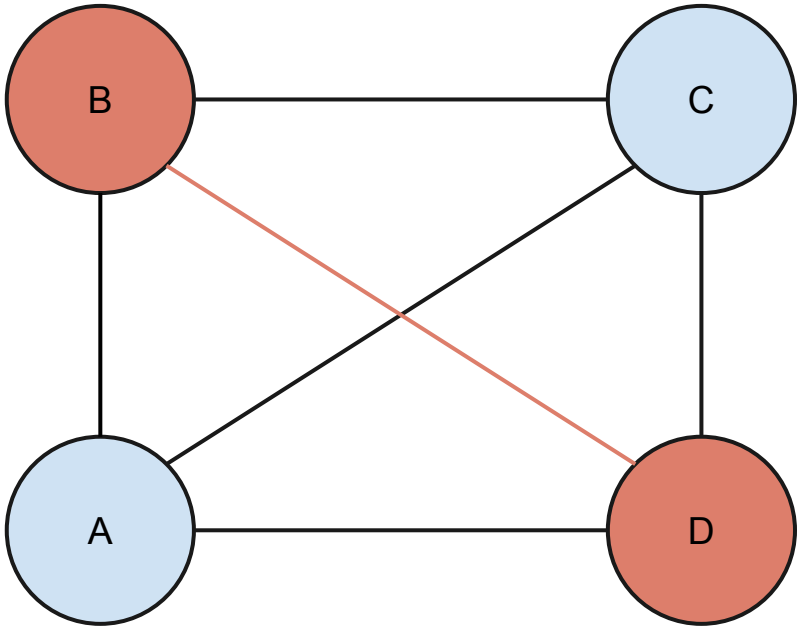
Algorithm 1: Skeleton

Graph $G \leftarrow$ complete undirected graph;
$l \leftarrow -1$;
while $l < |G|$ **do**
  $l \leftarrow l + 1$;
  foreach $\exists u, v \in G$ *s.t.* $|Adj(u) \setminus \{v\}| \geq l$ **do**
    if $v \in Adj(u)$ **then**
      foreach $k \subseteq Adj(u) \setminus \{v\}$ *s.t.* $|k| = l$ **do**
        if u, v *are conditionally independent given* k **then**
          remove edge $\{u, v\}$ from G;

PC

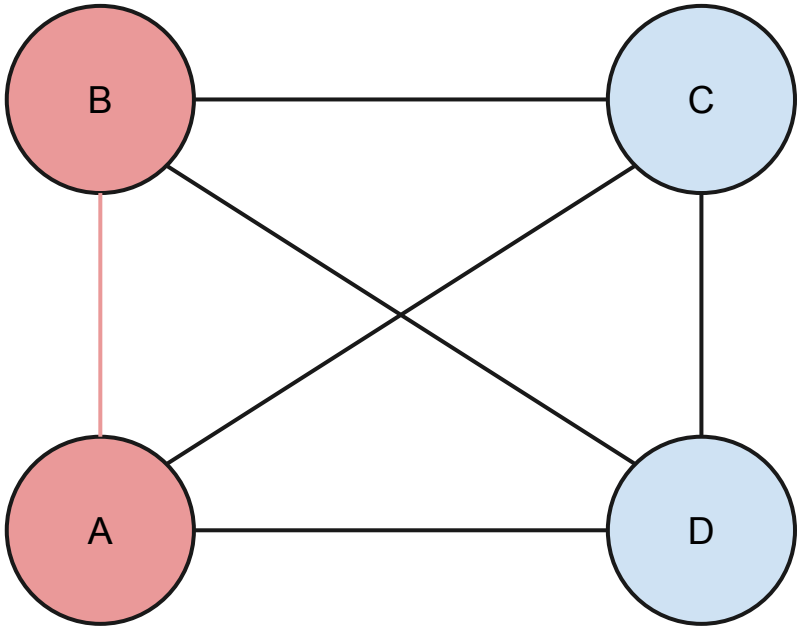
Dalla biologia sappiamo che i **geni** (nodi) B e D sono correlati

Vogliamo provare ad espandere questa rete con il PC



PC

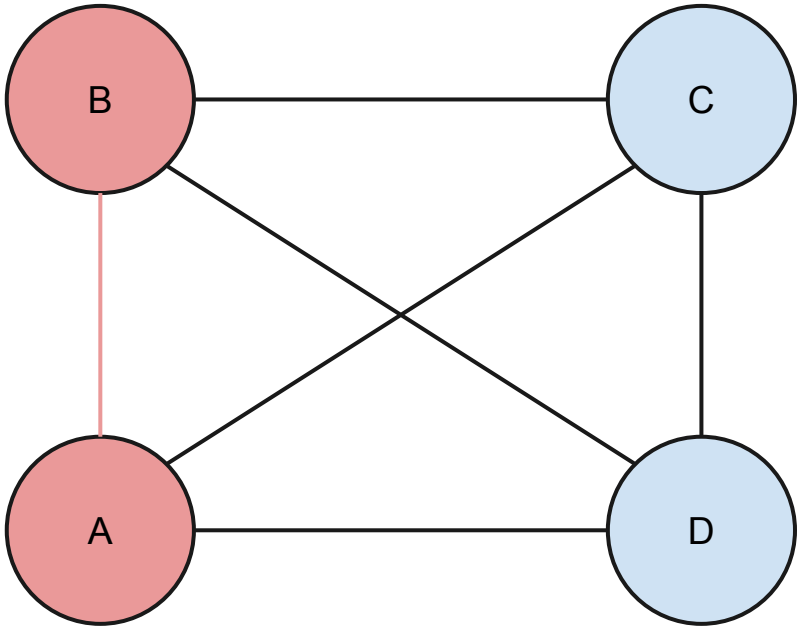
A e B sono indipendenti?



PC

A e B sono indipendenti?

Assumiamo siano poco correlati!

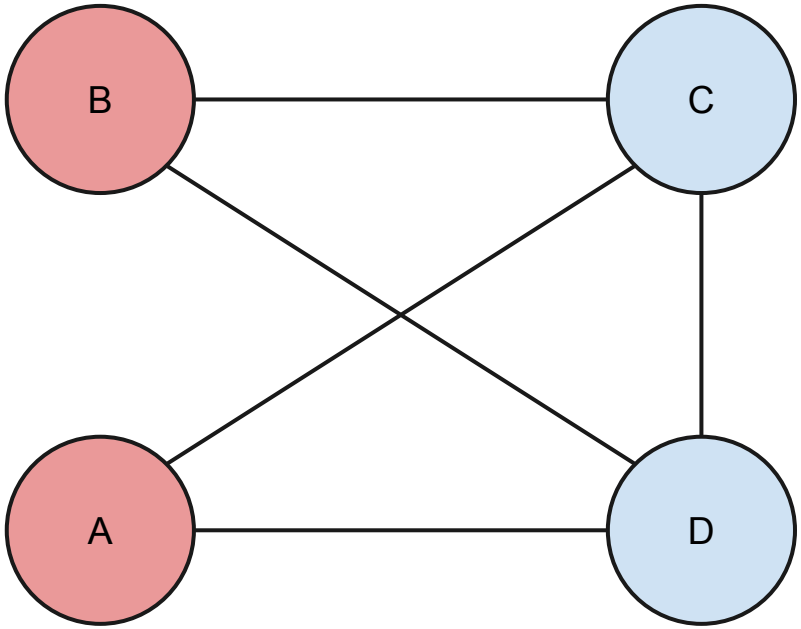


PC

A e B sono indipendenti?

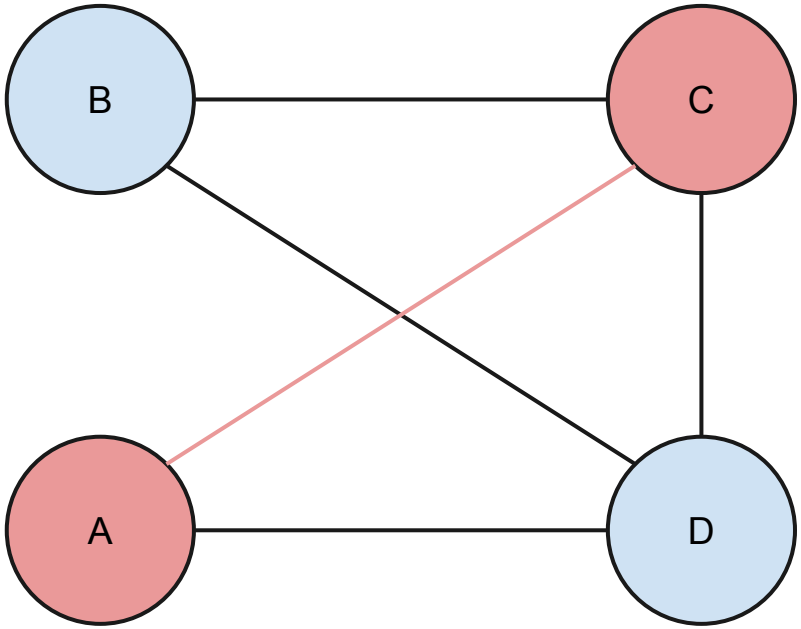
Assumiamo siano poco correlati!

Rimuoviamo l’arco tra A e B



PC

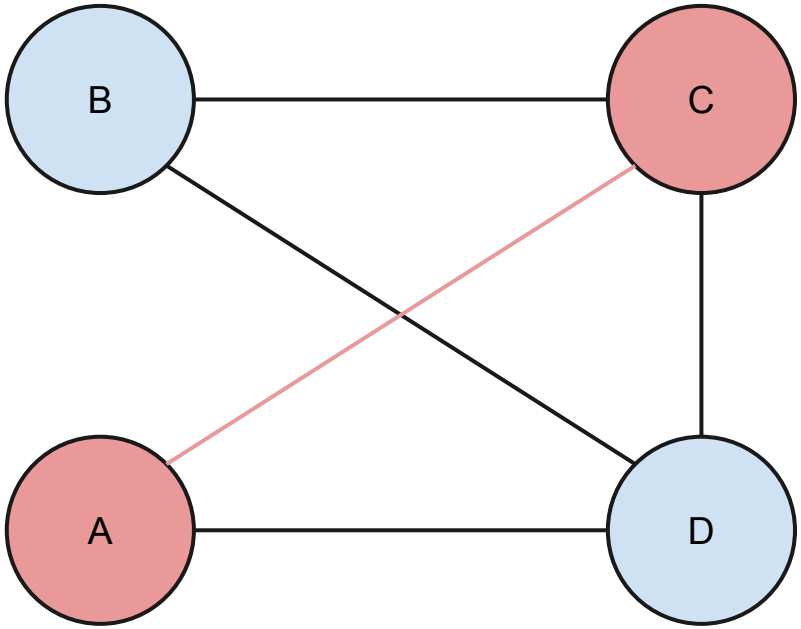
A e C sono indipendenti?



PC

A e C sono indipendenti?

Assumiamo siano poco correlati!

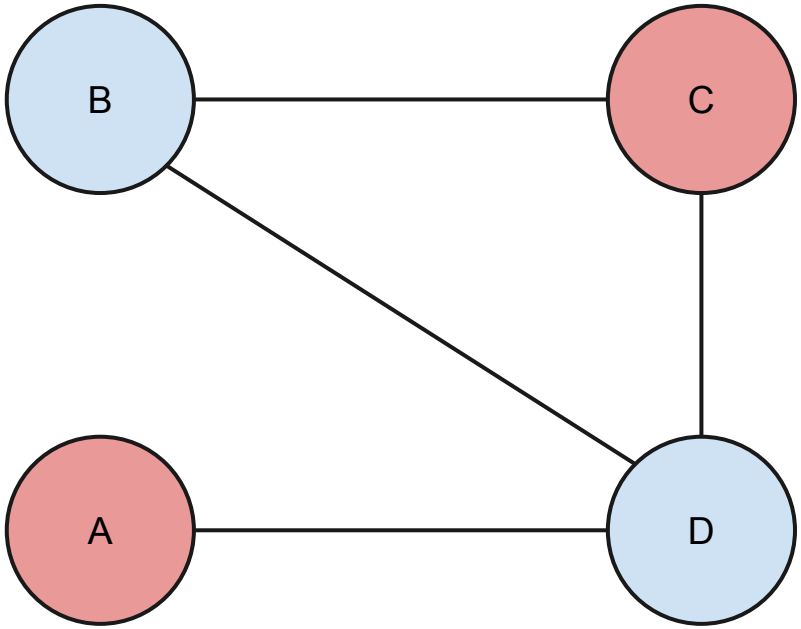


PC

A e C sono indipendenti?

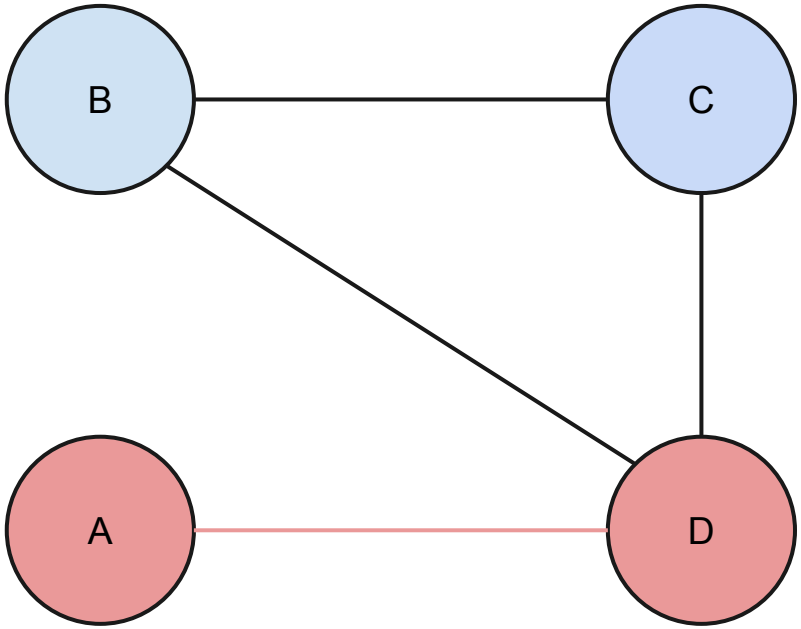
Assumiamo siano poco correlati!

Rimuoviamo l’arco tra A e C



PC

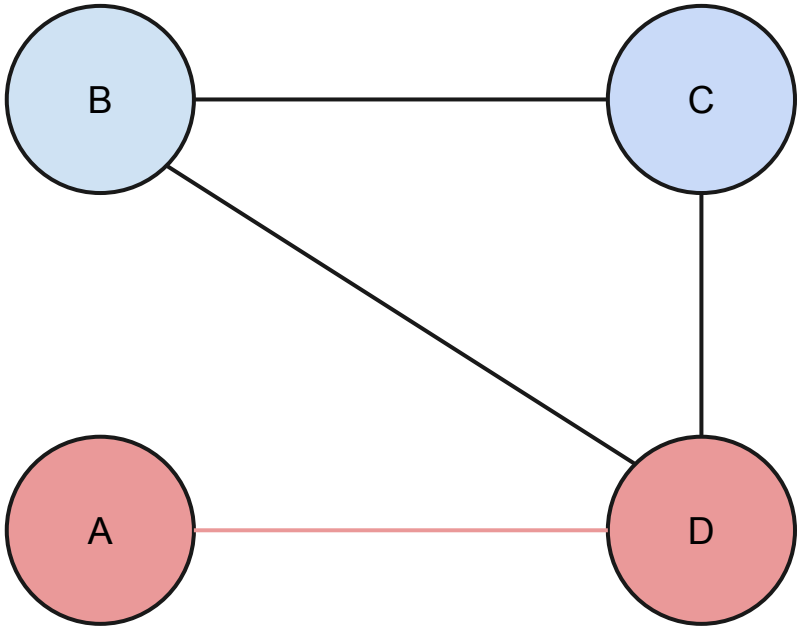
A e D sono indipendenti?



PC

A e D sono indipendenti?

Assumiamo siano poco correlati!

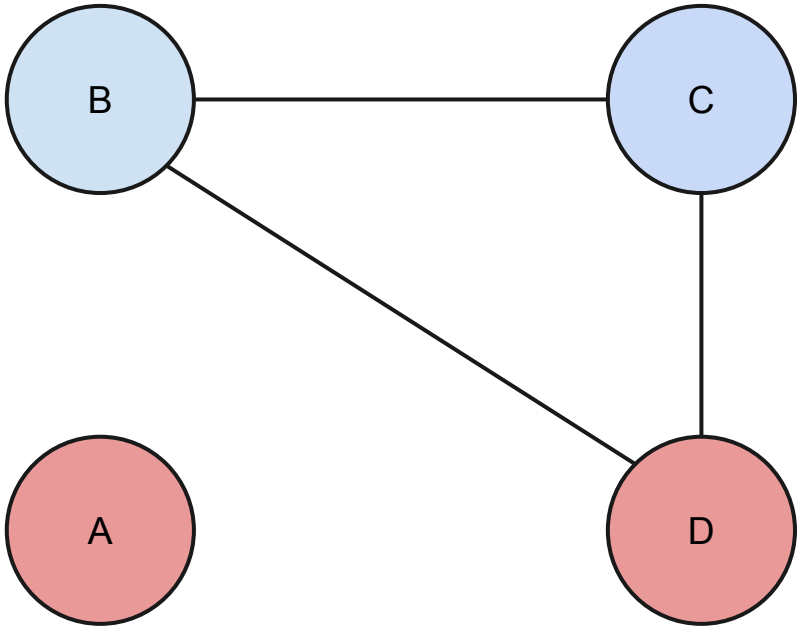


PC

A e D sono indipendenti?

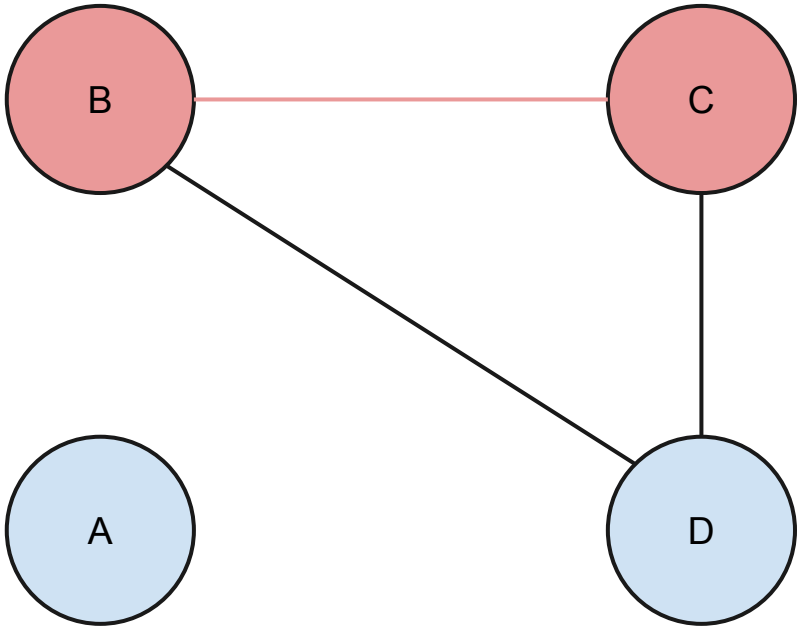
Assumiamo siano poco correlati!

Rimuoviamo l’arco tra A e D



PC

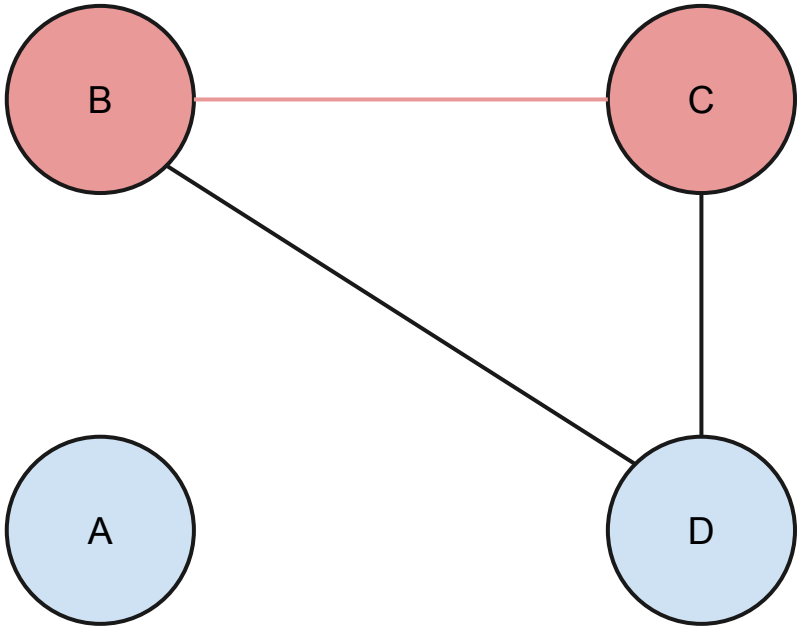
B e C sono indipendenti?



PC

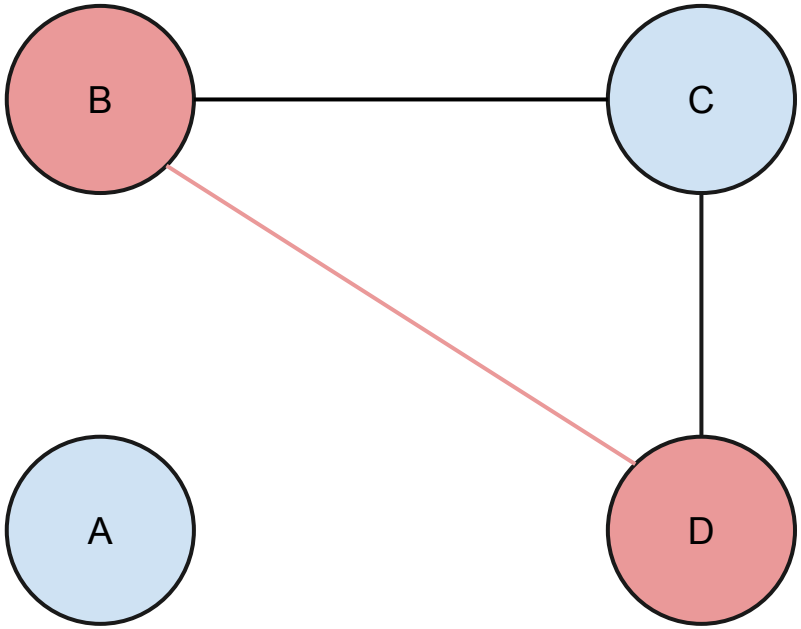
B e C sono indipendenti?

Assumiamo siano molto correlati!



PC

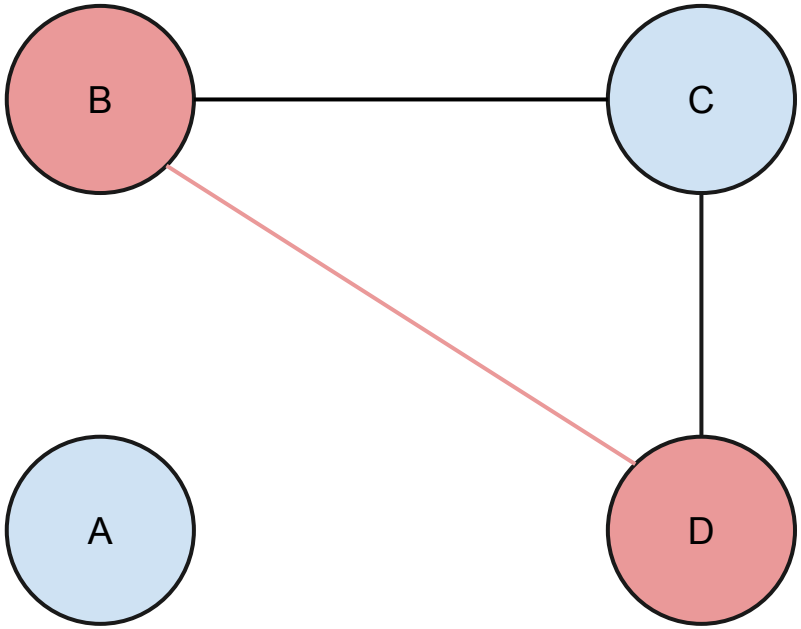
B e D sono indipendenti?



PC

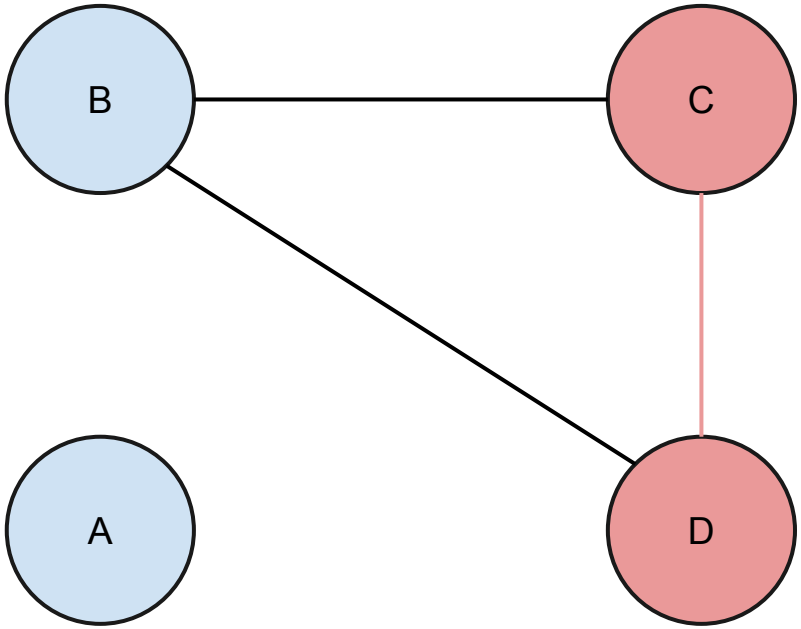
B e D sono indipendenti?

Assumiamo siano molto correlati!



PC

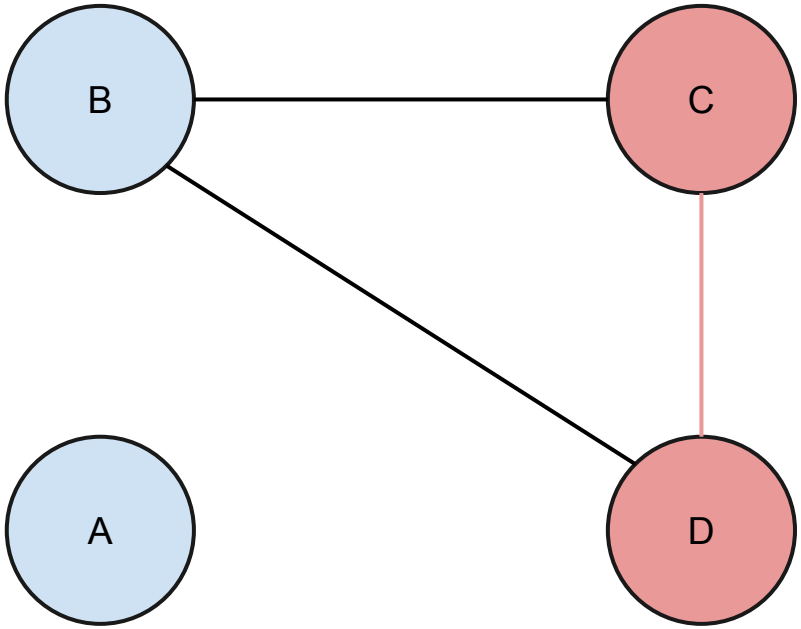
C e D sono indipendenti?



PC

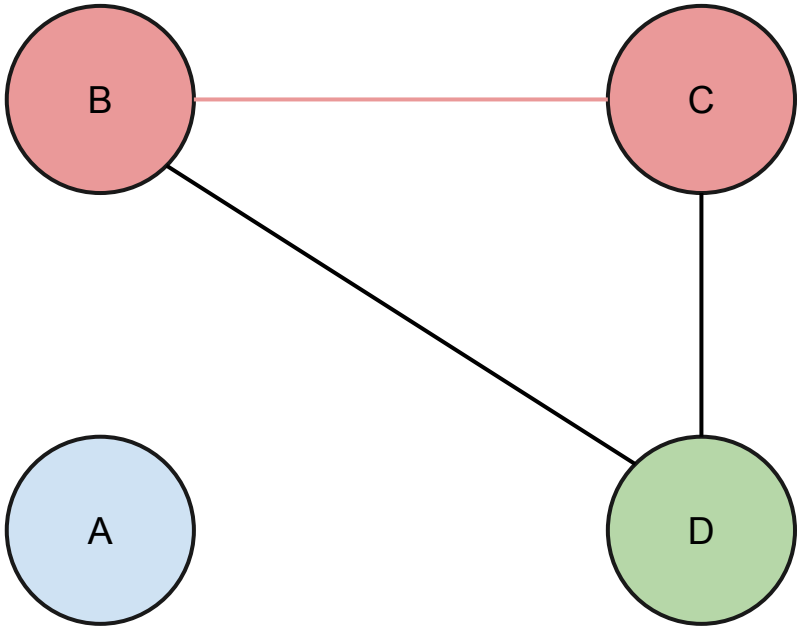
C e D sono indipendenti?

Assumiamo siano molto correlati!



PC

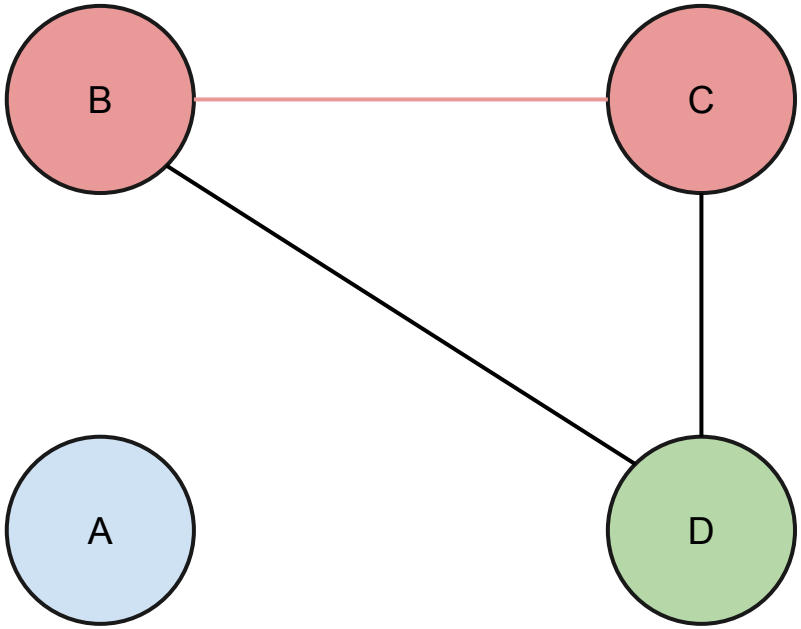
B e C dato D sono indipendenti?



PC

B e C dato D sono indipendenti?

Assumiamo diventino poco correlati a causa di D!

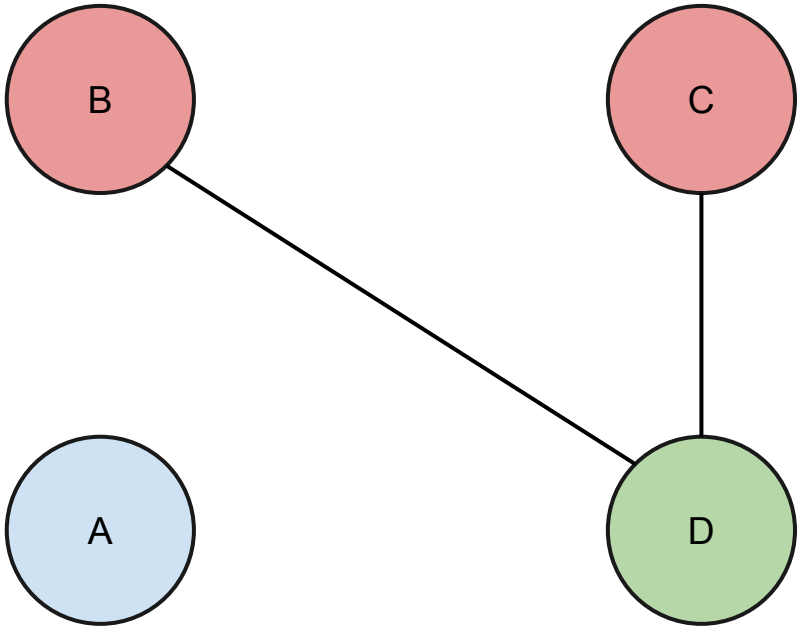


PC

B e C dato D sono indipendenti?

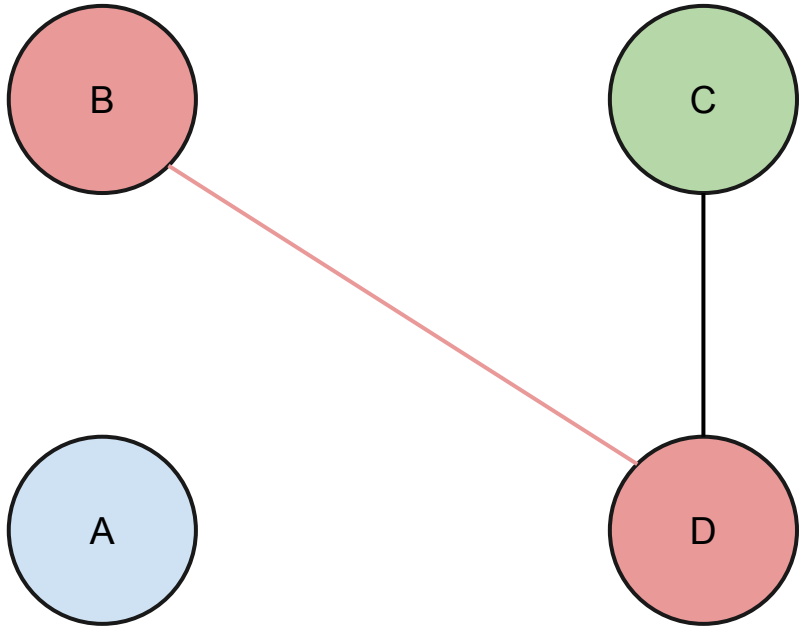
Assumiamo diventino poco correlati a causa di D!

Rimuoviamo l’arco tra B e C



PC

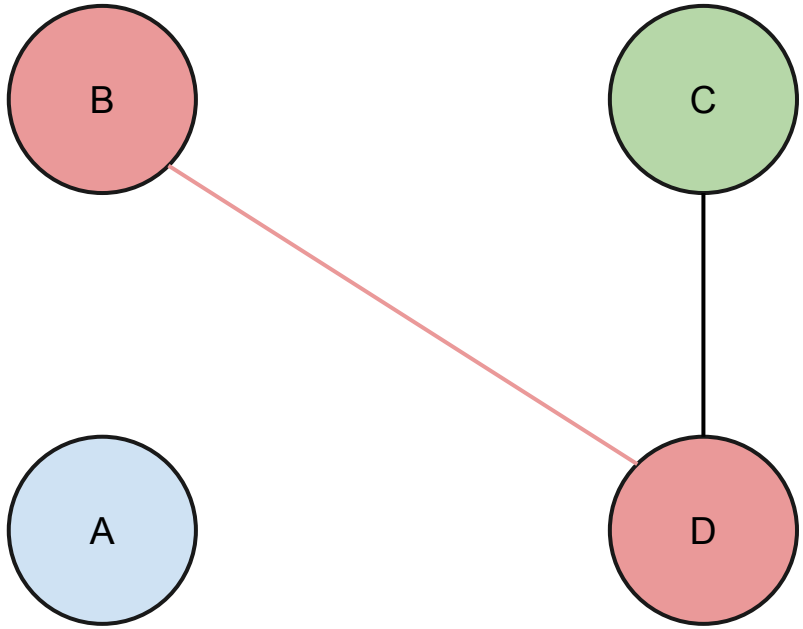
D e B dato C sono indipendenti?



PC

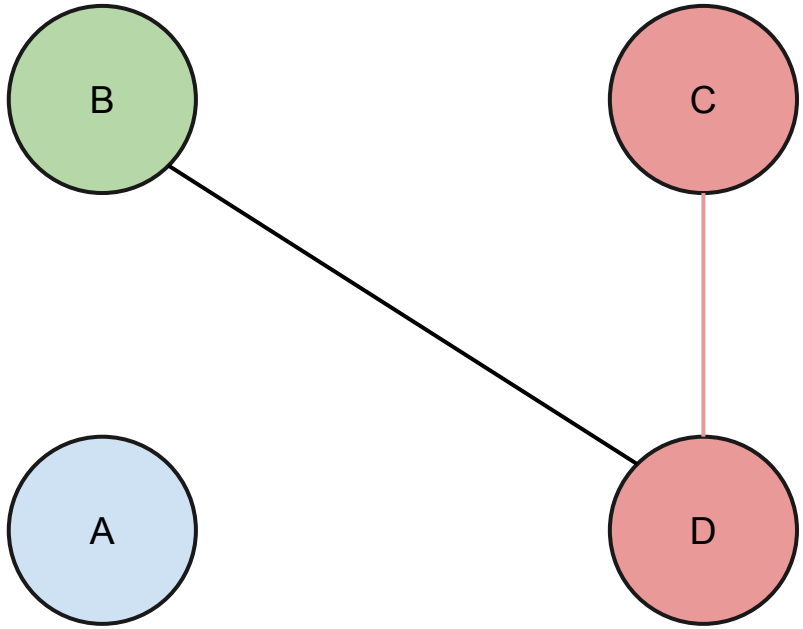
D e B dato C sono indipendenti?

Assumiamo restino molto correlati nonostante C!



PC

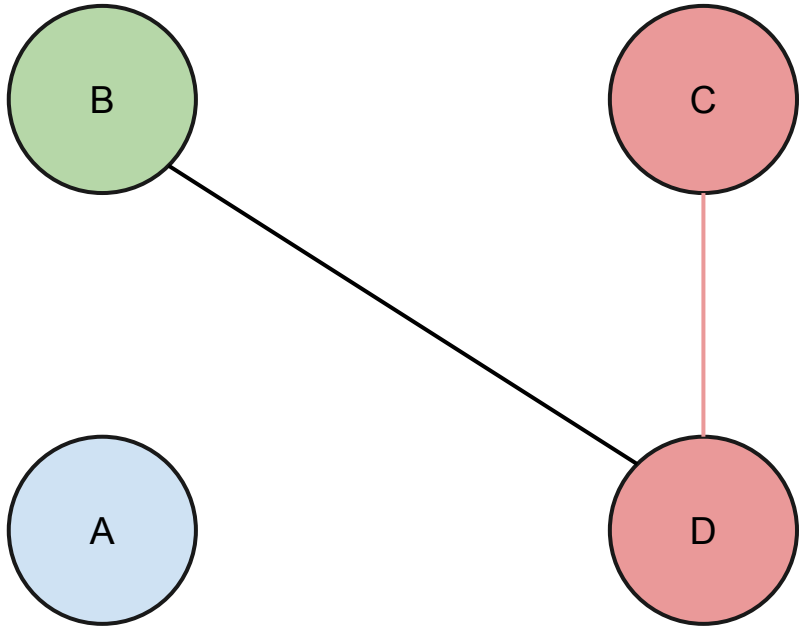
D e C dato B sono indipendenti?



PC

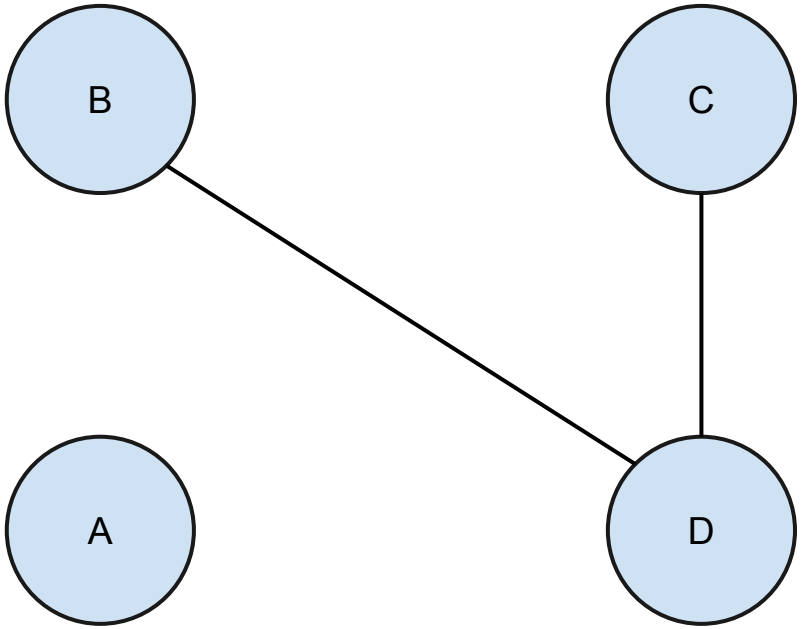
D e C dato B sono indipendenti?

Assumiamo restino molto correlati nonostante B!



PC

L’esecuzione del PC fornisce una nuova correlazione tra i nodi C-D, che è un **espansione** della rete genica di partenza (B-D)



PC

Avete notato l’**ordine** seguito nel considerare gli archi?

PC

Avete notato l'**ordine** seguito nel considerare gli archi?

Avremmo ottenuto un risultato diverso seguendo un ordine diverso?

PC

Avete notato l'**ordine** seguito nel considerare gli archi?

Avremmo ottenuto un risultato diverso seguendo un ordine diverso?

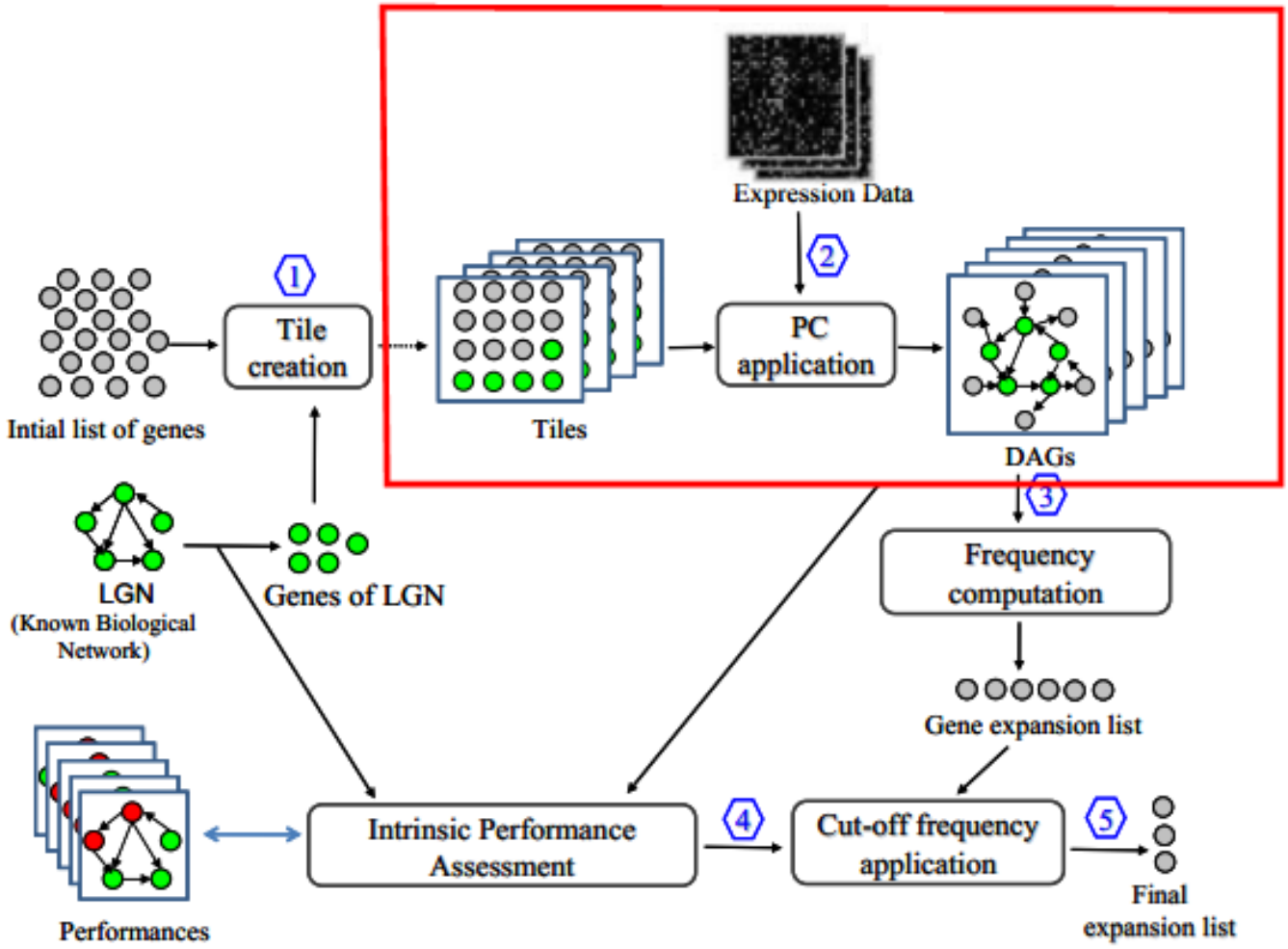
Purtroppo **SI**!

PC-IM

Il genoma può essere formato da decine di migliaia di geni, per esempio, il genoma dell'*Arabidopsis Thaliana* avrà un grafo con ~23.000 nodi (geni) e ~265.000.000 archi

Il **PC-IM** risolve il problema dell'ordine e della dimensione del genoma

PC-IM


Expression Data
1
Tile creation
2
PC application
Intial list of genes
Tiles
DAGs
3
Frequency computation
LGN
(Known Biological Network)
Genes of LGN
Gene expansion list
Intrinsic Performance Assessment
4
Cut-off frequency application
5
Final expansion list
Performances

PC-IM

- Tratta reti di 1000 nodi
- Servono 23 iterazioni per coprire un genoma
- Per fare analisi sulla frequenza servono almeno 100 iterazioni sul genoma
- Ogni PC-IM quindi consiste in 2.300 esecuzioni del PC
- Vogliamo espandere molte reti di un genoma, iterando l’espansione per diversi organismi

Outline

- Chi siamo
- gene@home
- BOINC
  - BOINC.Italy
- Implementazione
- Futuro

BOINC

B*erkeley* **O***pen* **I***nfrastructure*
for **N***etwork* **C***omputing*

Un framework open source per il calcolo volontario distribuito

BOINC

Supporta i sistemi operativi più comuni:

- Windows
- Mac OS X
- GNU/Linux
- FreeBSD e OpenBSD
- Solaris
- Android

BOINC

Siamo adatti per BOINC?

Attrazione sul Pubblico: il progetto deve essere recepito interessante e utile dai volontari

Basso rapporto tra dati e computazione: il traffico Internet e lo spazio su disco sono dei colli di bottiglia

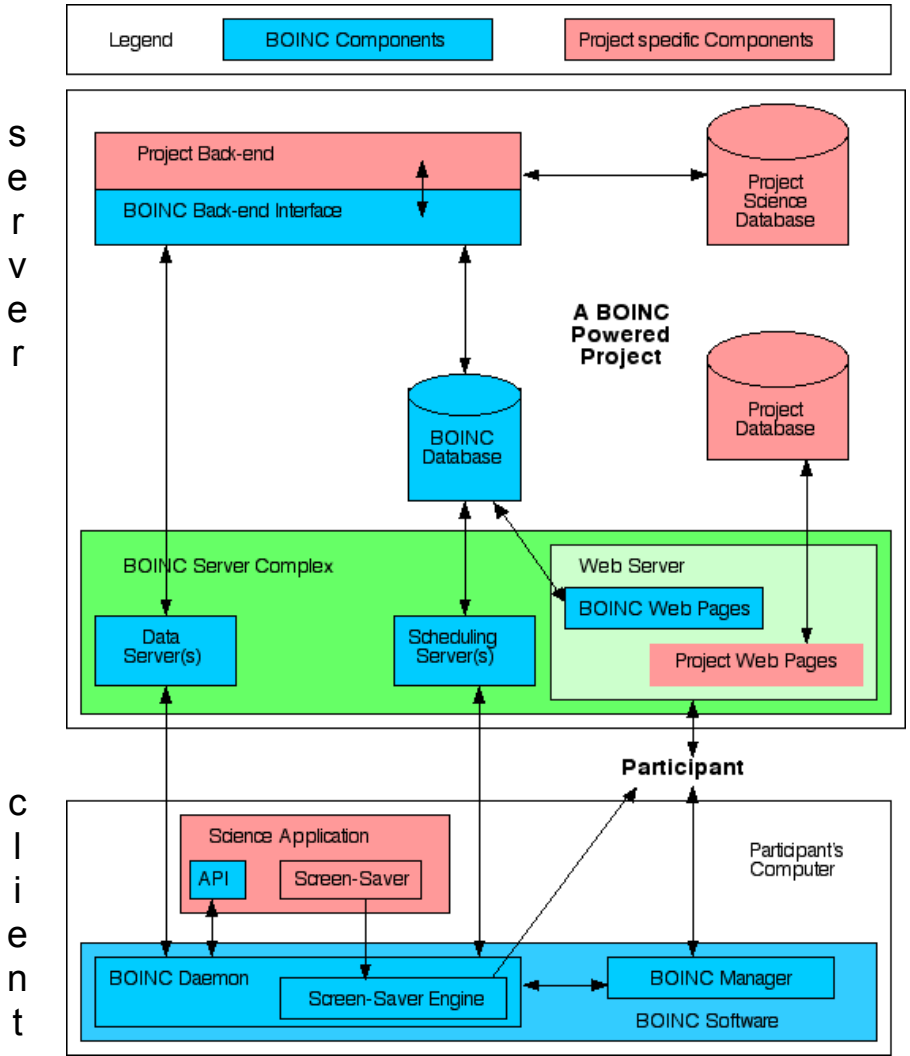
BOINC

Siamo adatti per BOINC?

Lavoro distribuito: poter suddividere un calcolo molto lungo in lavori più piccoli

Non-profit: il progetto deve essere *“di ricerca”* e non a fini di lucro

BOINC


Legend
BOINC Components
Project specific Components
server
Project Back-end
BOINC Back-end Interface
Project Science Database
A BOINC Powered Project
BOINC Database
Project Database
BOINC Server Complex
Web Server
BOINC Web Pages
Project Web Pages
Data Server(s)
Scheduling Server(s)
Participant
client
Science Application
API
Screen-Saver
Participant's Computer
BOINC Daemon
Screen-Saver Engine
BOINC Manager
BOINC Software

BOINC.Italy

Outline

- Chi siamo
- gene@home
- BOINC
- Implementazione
  - C++ = D
  - Grafo
  - Programmazione Dinamica
  - Performance
- Futuro

C++ = D

PRO

- Velocità
- Gestione della memoria dinamica
- Nativo per l’integrazione con le API di BOINC
- Compilatore disponibile per tutte le piattaforme

C++ = D

CONTRO

- Responsabilità al programmatore nella gestione della memoria allocata
- Si deve compilare l'applicazione per ogni tipo di architettura presente (per esempio, 32 e 64 bit)
- Alcune piattaforme utilizzano librerie proprietarie con funzioni non standard

C++ = D

Sistemi Operativi supportati:

- Windows (32/64 bit) da XP
- Mac OS X (CPU Intel) dalla 10.5
- GNU/Linux (32/64 bit) con kernel 3.x

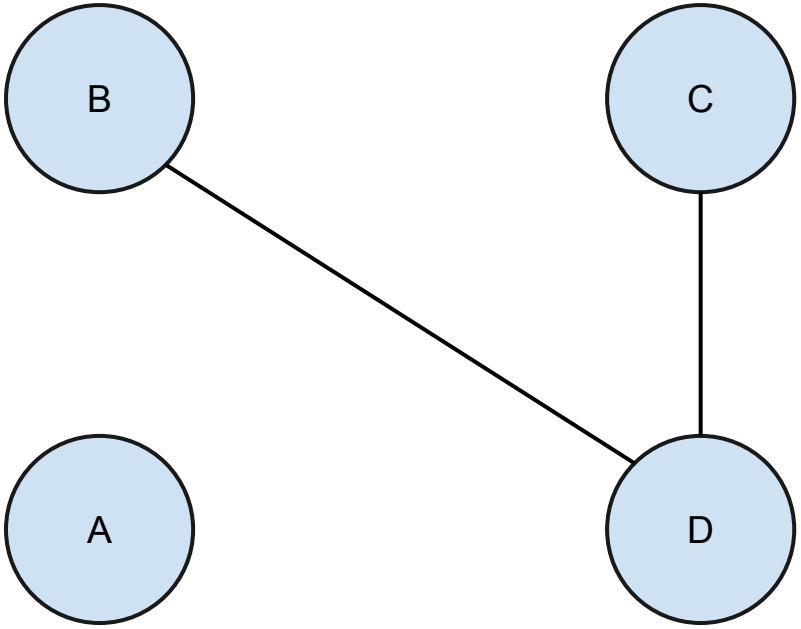
C++ = D

Un esecuzione può durare diverse ore.

Attraverso un sistema di **checkpoint**, è possibile sospendere l'esecuzione e riprenderla in un qualsiasi altro momento, senza perdita del lavoro compiuto

Grafo

Abbiamo utilizzato una **matrice booleana** (chiamata anche "*matrice di adiacenza*") per rappresentare il grafo della rete genica



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

Programmazione Dinamica

$$\rho_{i,j|k} = \frac{\rho_{i,j|k\setminus h} - \rho_{i,h|k\setminus h}\rho_{j,h|k\setminus h}}{\sqrt{(1-\rho^2_{i,h|k\setminus h})(1-\rho^2_{j,h|k\setminus h})}}$$

$O(3^{\ell})$

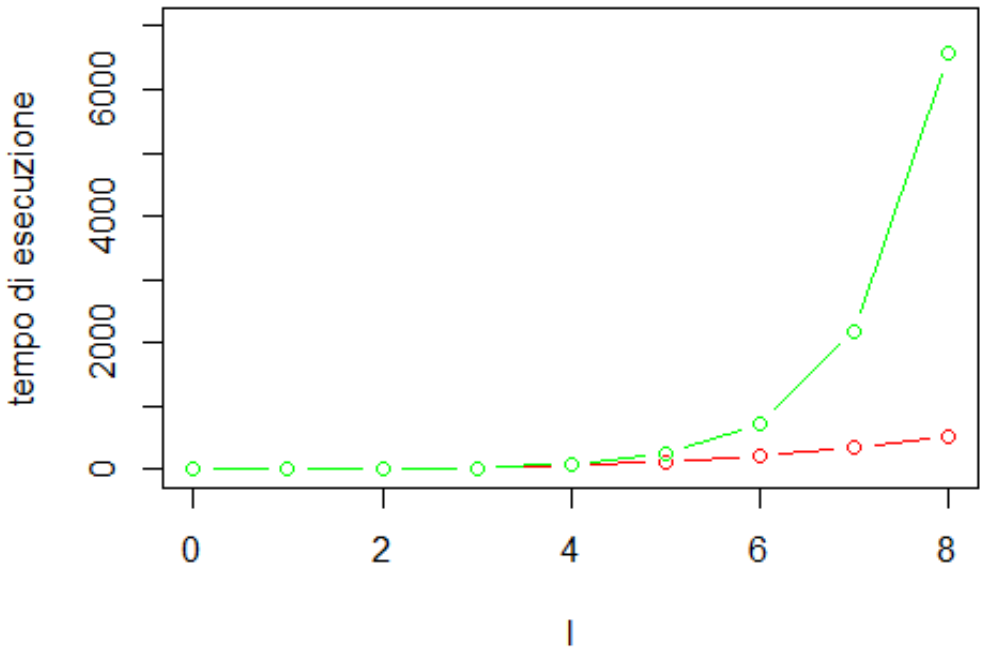
Algorithm 2: Correlation

function *Dynamic correlation (int l, matrix ρ)*

```
dim ← l + 2;
for k = 1 to l do
    for i = 0 to l − k do
        for j = i + 1 to dim − k do
            ρ[i][j] = ρ[j][i] = (ρ[i][j]−ρ[i][dim−k]∗ρ[j][dim−k]) / √((1−ρ²[i][dim−k])∗(1−ρ²[j][dim−k],2));
return ρ[0][1];
```

$O(\ell^3)$

Programmazione Dinamica



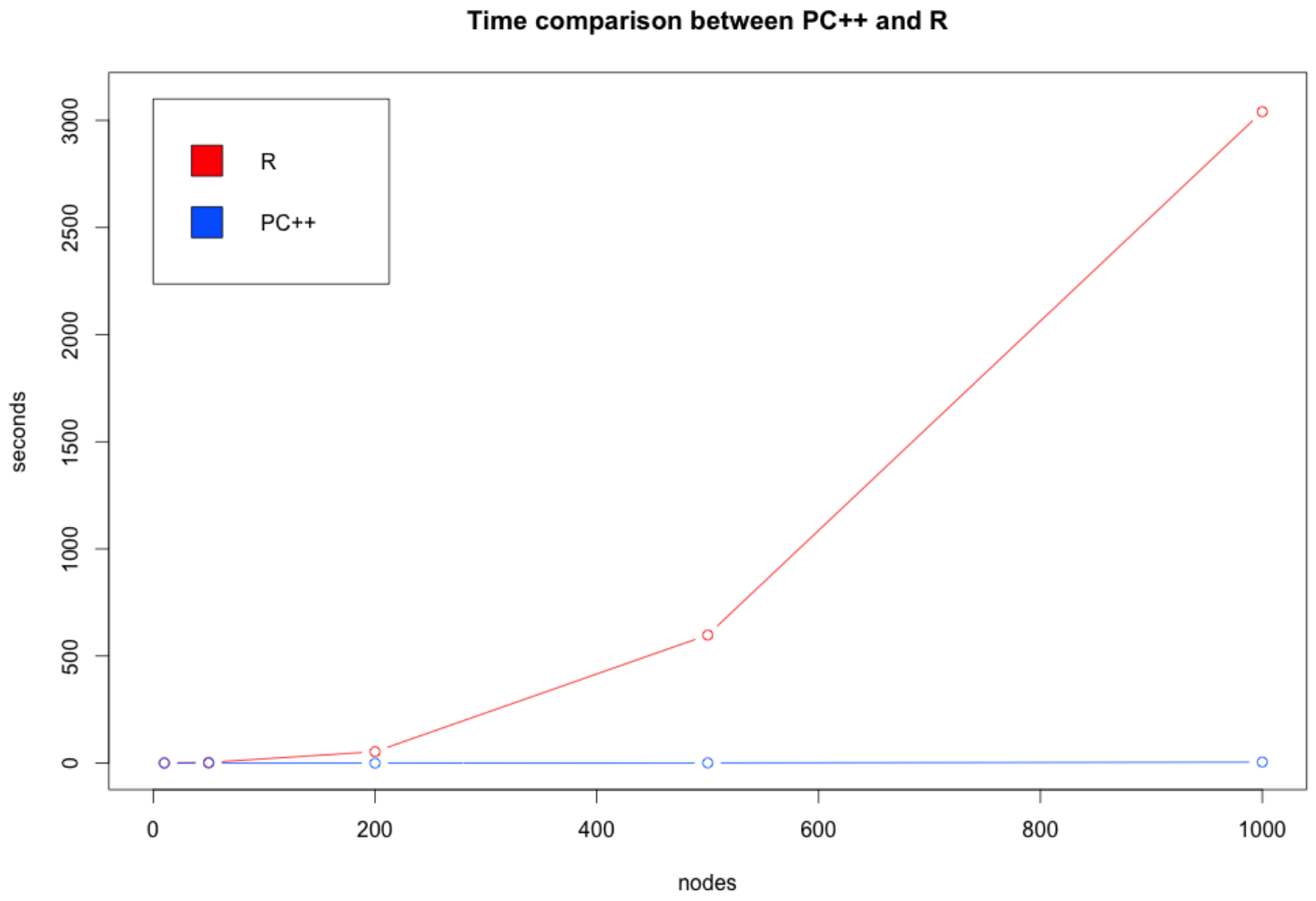
$O(3^{\ell})$ $O(\ell^3)$

La programmazione dinamica è una tecnica che evita, in caso di ricorsione, il calcolo di valori già precedentemente computati

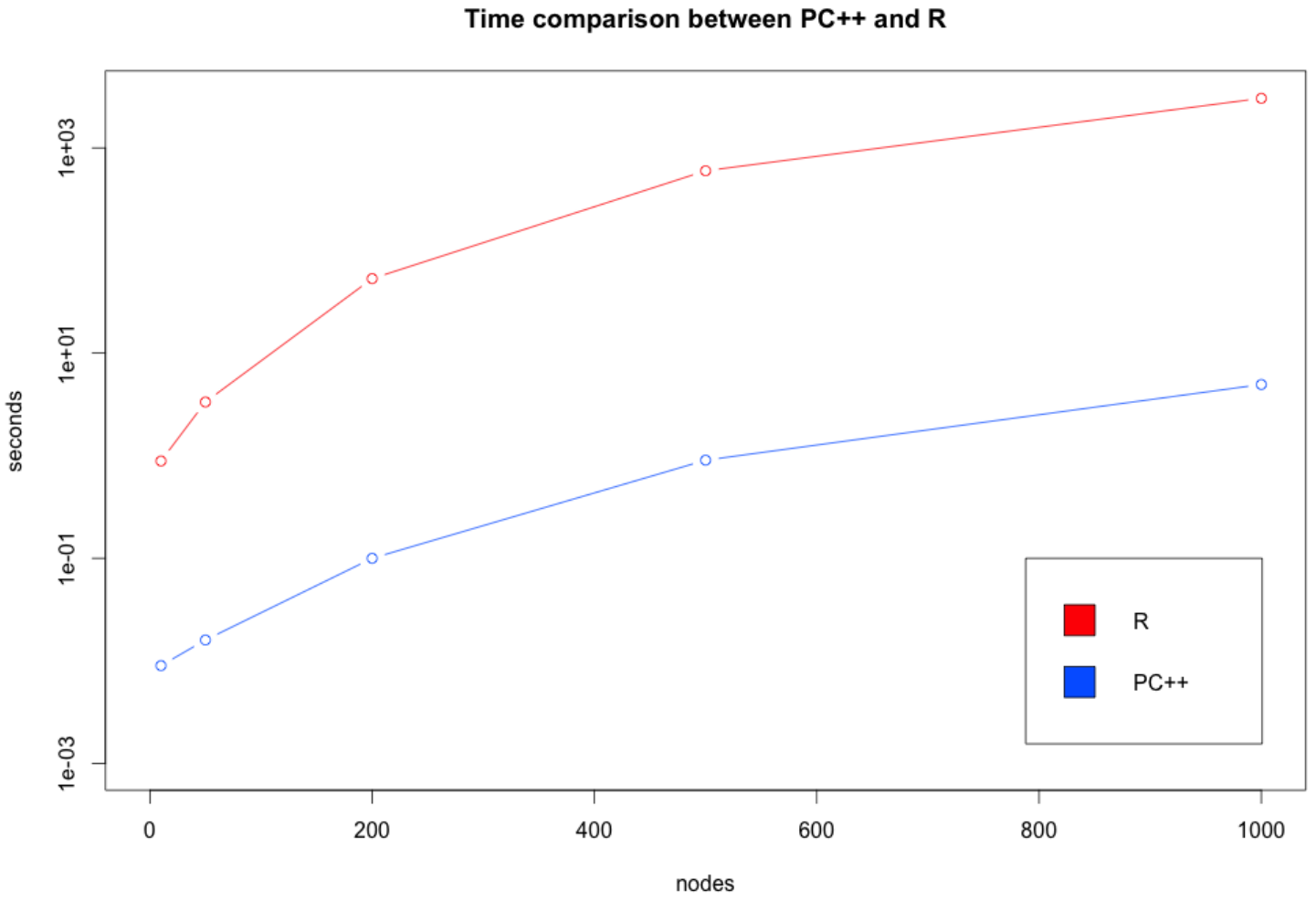
Performance

Esiste una implementazione del **PC** in R (ambiente di sviluppo per analisi statistica) che non è efficiente né dal punto di vista del tempo di esecuzione, né dalla memoria

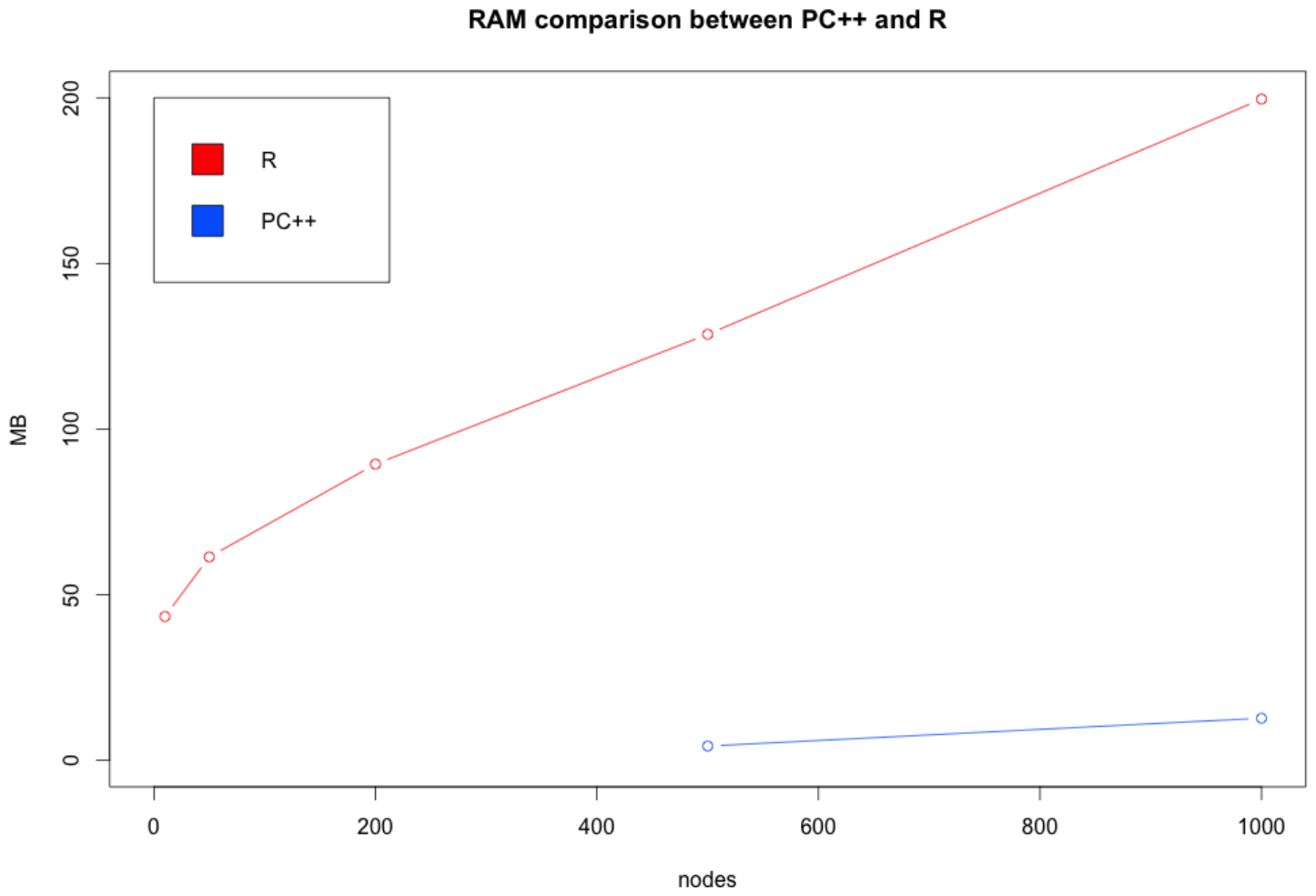
Performance



Performance



Performance



Outline

- Chi siamo
- gene@home
- BOINC
- Implementazione
- Futuro

Futuro

Gestione del server BOINC per supportare il progetto con la **generazione** di nuove reti genetiche da espandere

Bug-fixing delle diverse applicazioni disponibili per le varie piattaforme

Analisi dei risultati computati da BOINC per poter valutare l'espansioni delle reti in biologia

Futuro

Pubblicazione **open source** della nostra implementazione del **PC**, in collaborazione con l'*Università degli Studi di Trento*

Grazie per l’attenzione

Domande?