

Network Alignment

Studenti:

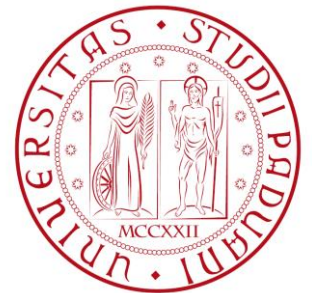
Luca Masiero

Stefano Ivancich



Supervisor:

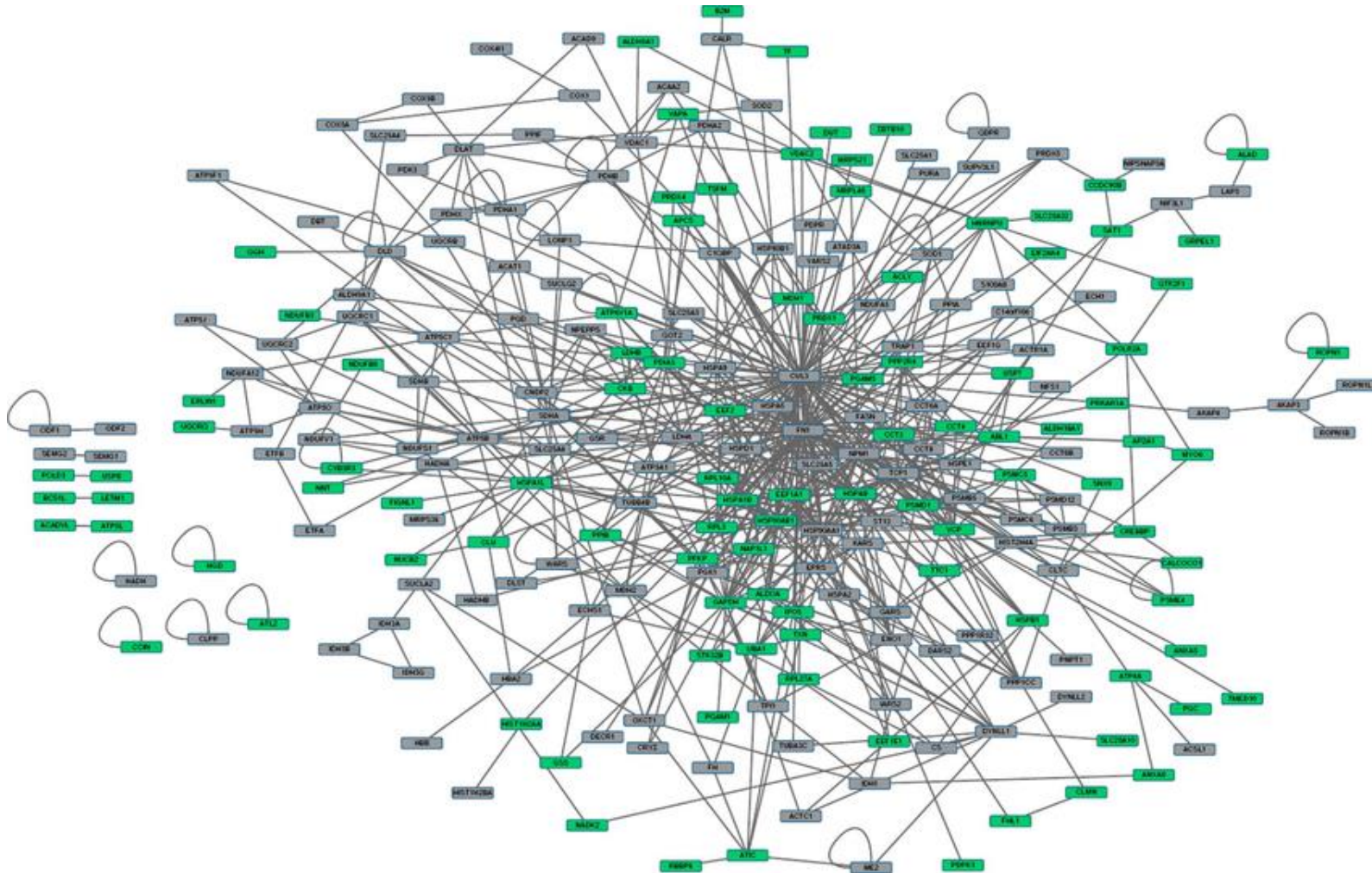
Prof. Matteo Comin



11 Giugno 2020

Introduzione: *Network Alignment* e PPIN

Obiettivo = trovare somiglianze tra struttura e/o topologia di due o più reti.



Le *Protein-Protein Interaction Networks* (PPIN) sono strumenti validi per comprendere:

- *funzioni* delle cellule;
- *malattie umane*;
- design e riposizionamento dei *farmaci*;
- *interattomi* (insieme delle interazioni molecolari in una cellula).

Date le grandi dimensioni (migliaia di elementi), le reti PPI sono analizzate tramite l'identificazione di **moduli**.

Modulo topologico = gruppo di nodi che hanno molte più connessioni con i nodi del gruppo piuttosto che con quelli esterni.

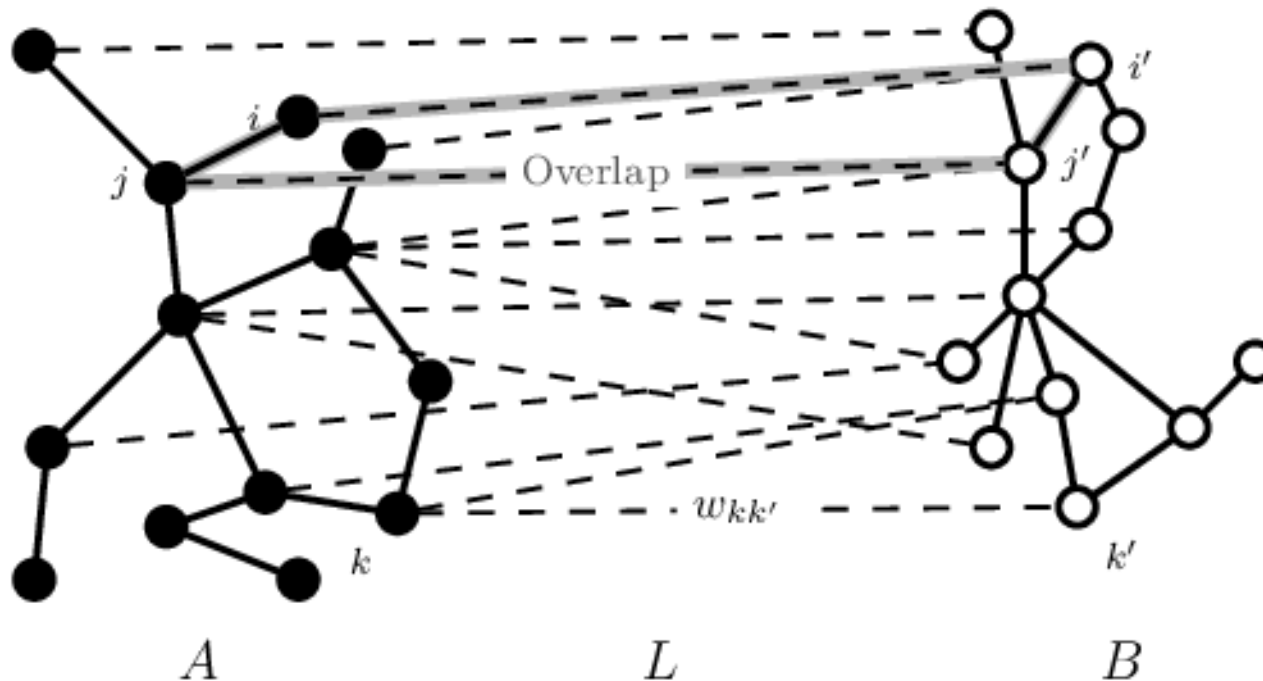
Modulo funzionale = gruppo di nodi che condividono una funzione biologica.

Introduzione: *Network Alignment* e PPIN

Date due reti, **allinearle** significa trovare un *mapping* in grado di:

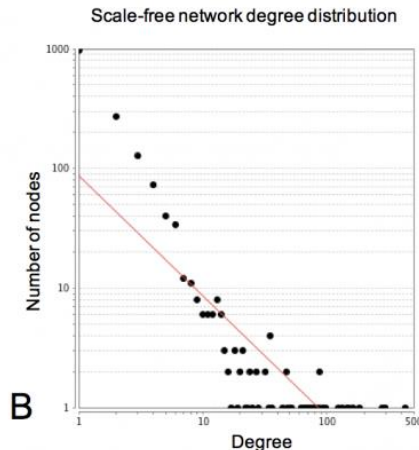
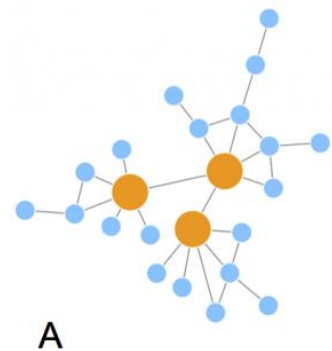
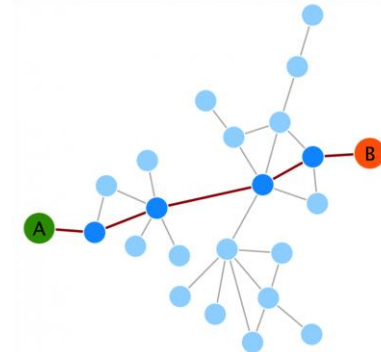
- (1) *massimizzare il numero di proteine mappate (nodi)* che sono correlate da un punto di vista funzionale;
- (2) *massimizzare il numero di interazioni comuni (archi)* tra le reti.

Problema intrattabile dovuto all'*NP-completezza* del *sub-graph isomorphism problem* (Cook, 1971).



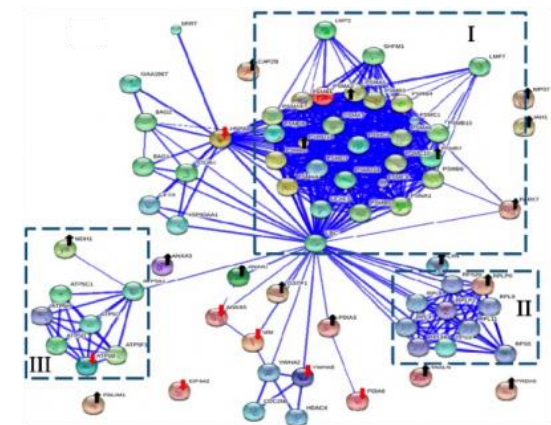
PPIN: proprietà fondamentali

Effetto del piccolo mondo: tutte le reti complesse sono tali che due nodi qualsiasi possono essere collegati da un percorso costituito da un numero relativamente piccolo di collegamenti.

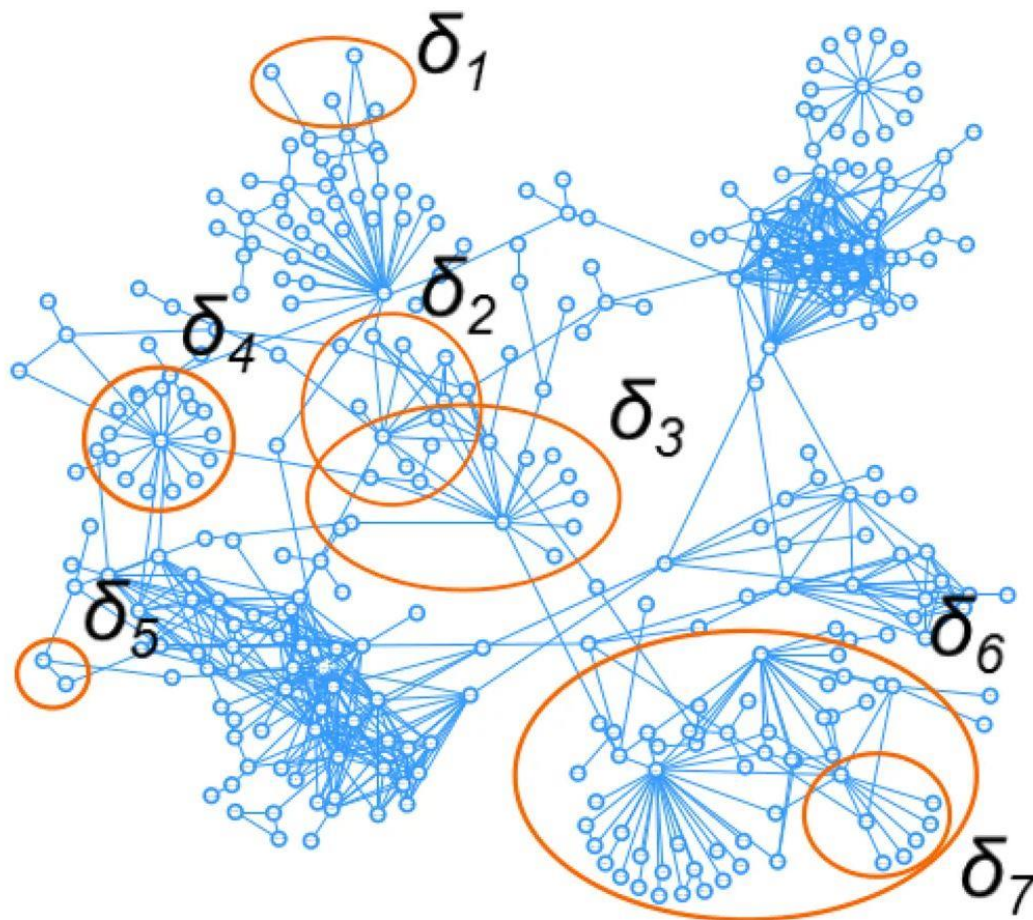


Scale-free networks: nodi con poche connessioni vs *hub*

PAROLE CHIAVE scalabilità, invarianza ai cambiamenti di scala, vulnerabilità agli attacchi mirati.



Transitività: misura la tendenza dei nodi a raggrupparsi. Utile per individuare *complessi proteici (moduli)*.



Metodo **GO-based** per
identificare moduli
funzionali.

Combinazione di
informazioni provenienti
dalla **topologia** delle reti con
conoscenza biologica.

Overlapping e **copertura
totale** della rete.

3 FASI

1) Inizializzazione: creazione delle partizioni e dei moduli

$$C = \{c_1, \dots, c_h, \dots, c_H\}$$

$$c_1 \cap c_2 \dots \cap c_h \dots \cap c_H \equiv \emptyset$$

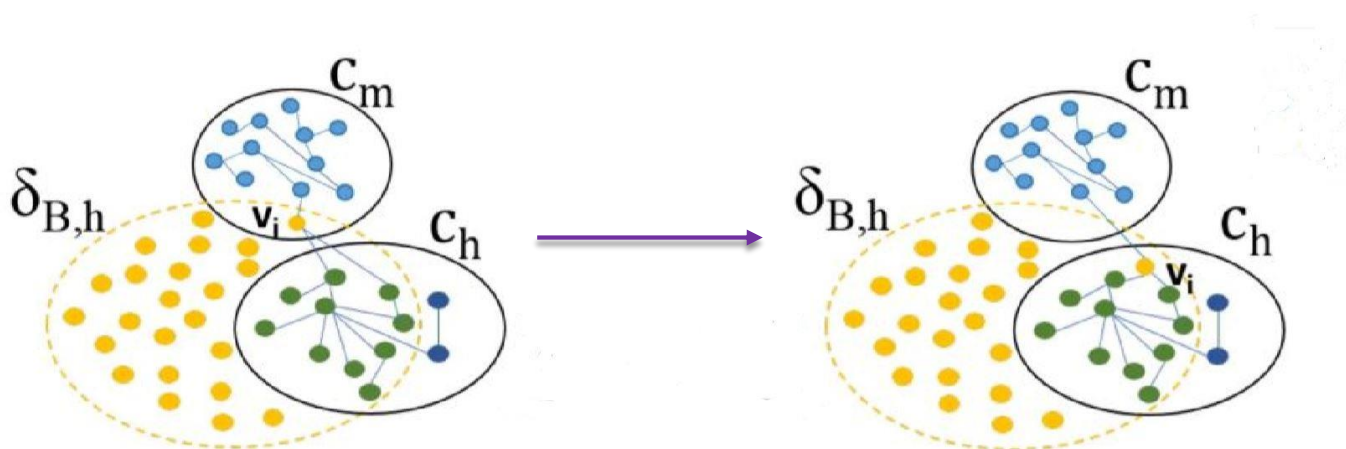
$$c_1 \cup c_2 \dots \cup c_h \dots \cup c_H \equiv V$$

$$\Phi = \{\phi_1, \dots, \phi_h, \dots, \phi_H\}$$

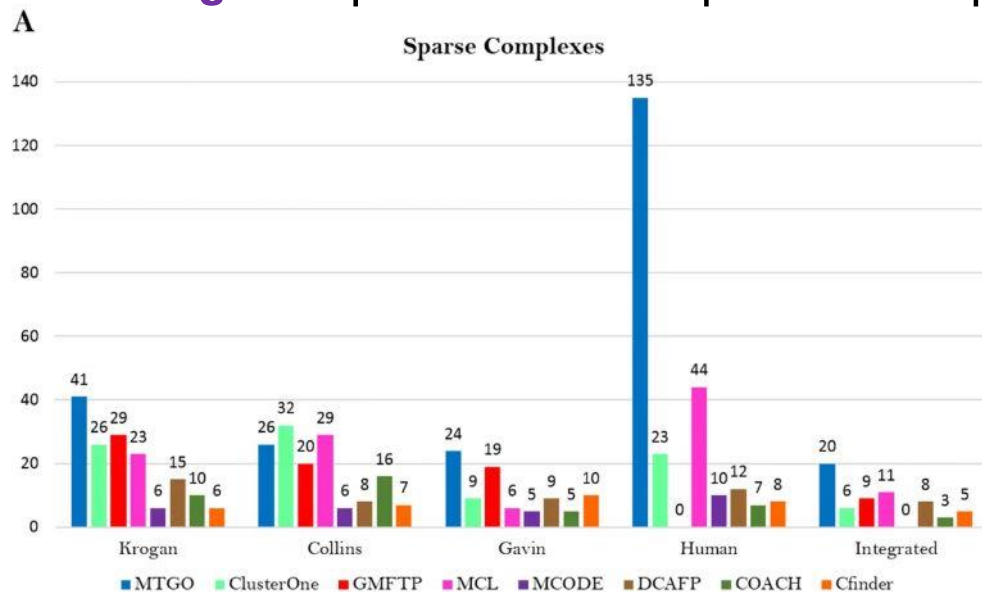
$$\phi_1 \cap \phi_2 \dots \cap \phi_h \dots \cap \phi_H \equiv \emptyset$$

$$\phi_1 \cup \phi_2 \dots \cup \phi_h \dots \cup \phi_H \subseteq V$$

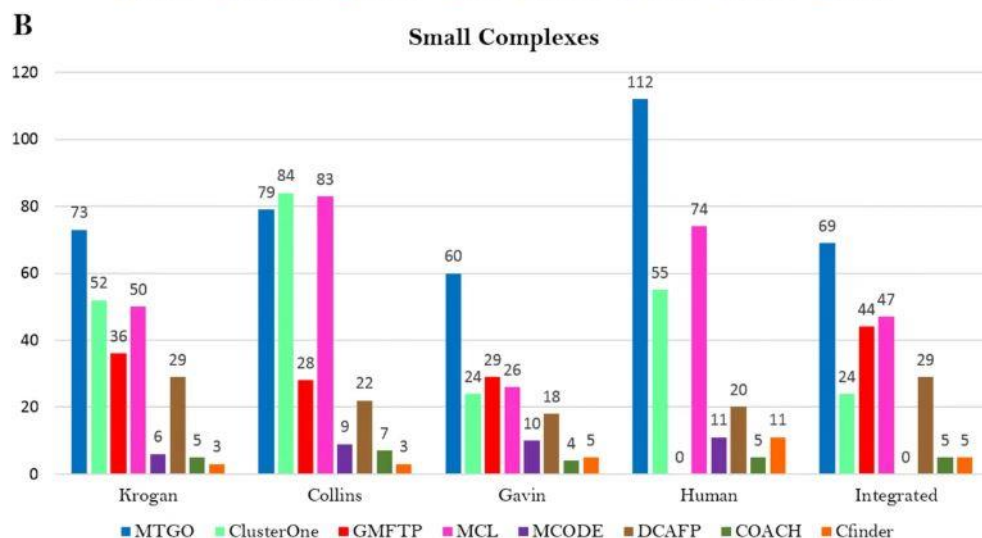
2) Iterazioni in cui i nodi vengono ri-assegnati alle partizioni



3) **Convergenza** per valutare la qualità della partizione finale e dell'overlapping



MTGO possiede l'abilità di individuare un insieme di **termini GO** fornendo un'interpretazione biologica significativa della PPIN (proprietà assente negli altri algoritmi allo stato dell'arte).



NUMERO DI CITAZIONI = 13



Metodo per l'**allineamento di più PPIN**.

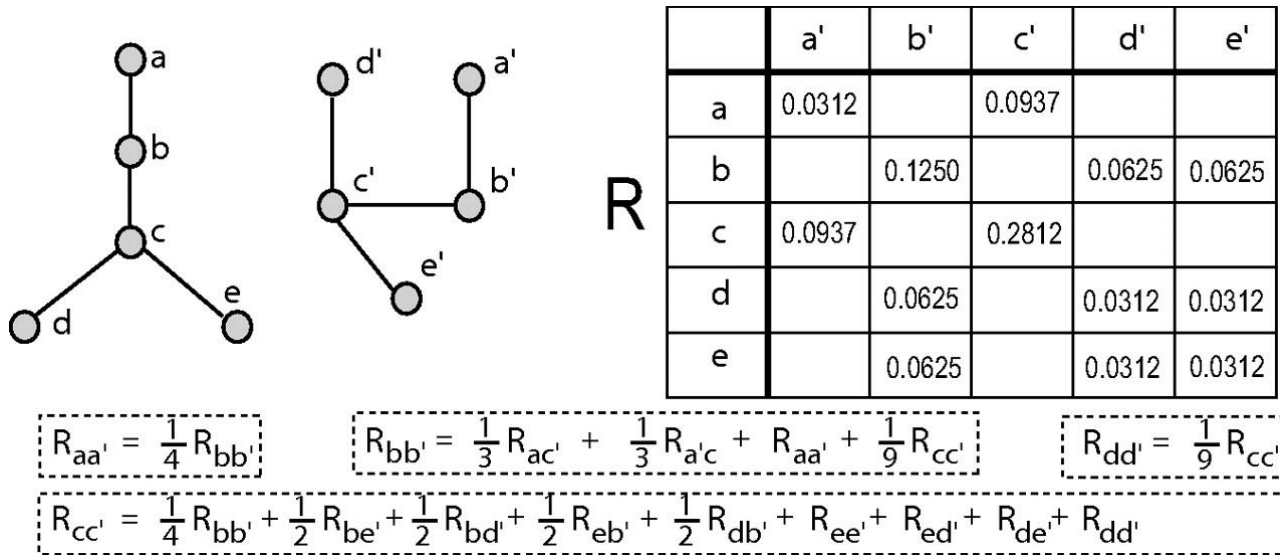
Intuizione = una proteina rappresenta una buona corrispondenza con una proteina in un'altra sequenza se le rispettive sequenze e i loro intorni topologici costituiscono una buona corrispondenza.

Rappresenta un *approccio di analisi comparativa* al GNA.

Caso di GNA a coppie

Input: due PPIN G_1 e G_2 , ogni arco e può aver associato un peso $w(e)$ ($0 \leq w(e) \leq 1$) e *similarity measure* tra i nodi delle due reti.

IsoRank: Fasi



Obiettivo: trovare **sottografo** comune alle reti in input.

2 FASI:

- 1) **Assegnazione** di *functional similarity scores*;
- 2) **Mapping** per il GNA (solo scores elevati) mantenendo la proprietà transitiva
(**one-to-one** o **many-to-many**)

$$R = \sum R_{ij} = \sum_{u \in N(i)} \sum_{v \in N(j)} \frac{1}{|N(i)||N(j)|} R_{ij} \text{ con } i \in V_1, j \in V_2$$

Il sottografo corrispondente all'allineamento globale possiede 1663 archi in comune ad almeno due PPIN e 157 archi in comune al almeno 3 PPIN.

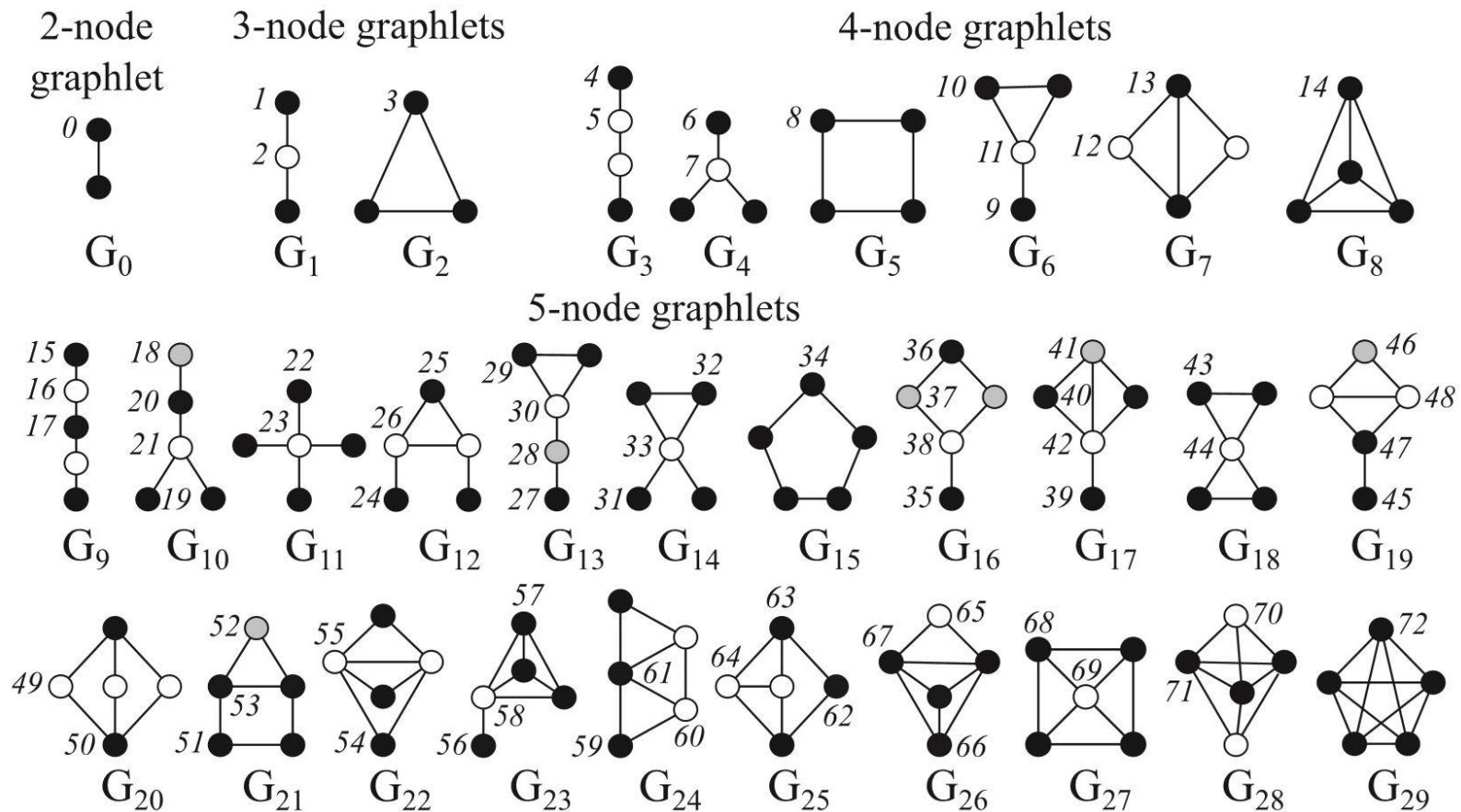
La dimensione del sottografo comune relativamente piccola (*overlap* con $\approx 5\%$ della PPIN umana) a causa dell'incompletezza e della rumorosità dei dati.

All'aumentare della quantità e della qualità dei dati, l'*overlap* dovrebbe aumentare. Delle 86932 proteine provenienti dalle 5 specie, 59539 (68,5%) hanno ottenuto almeno un match con un'altra proteina di una rete diversa.

NUMERO DI CITAZIONI = 505 dallo sviluppo nel 2008.

 Google Scholar

Idea: mappare insieme nodi che costituiscono un pattern (*graphlet*) con molte interazioni condivise.



Funzione obiettivo: fonde le informazioni dalle sequenze di proteine con le interazioni tra i *graphlet* (si risolve con la **Programmazione Intera**).

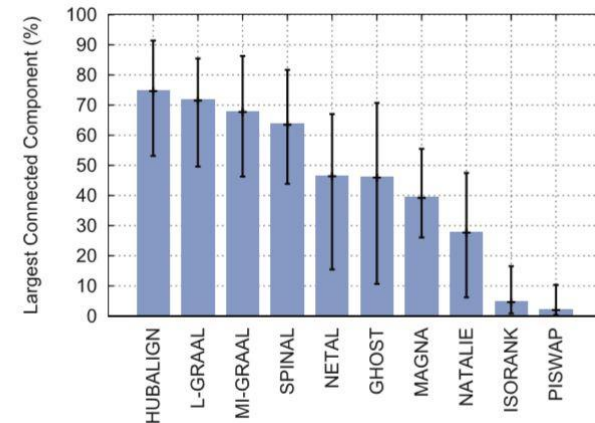
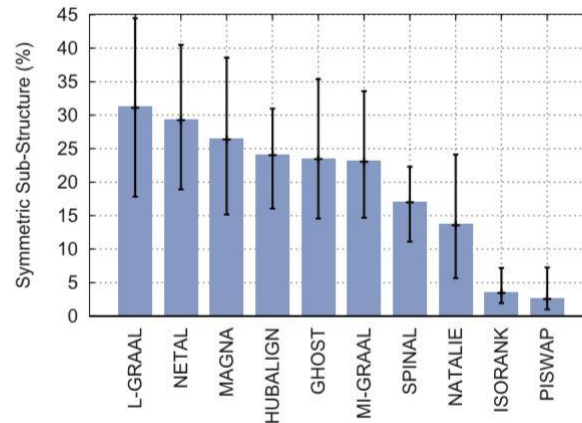
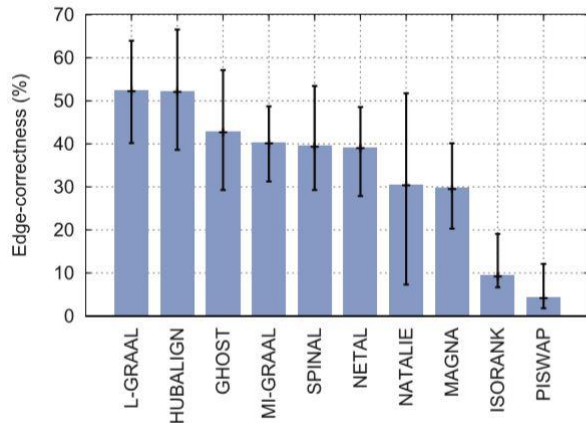
FASI

- 1) *Similarity scores* e definizione della *topological similarities*;
- 2) Risoluzione dell'equazione

$$IP = \max_{x,y} \left(\alpha \sum n(i, k) \times x_{ik} + (1 - \alpha) \sum e(i, j, k, l) \times y_{ijkl} \right)$$

in tempo $O(|V|^3 + |V|^2 \times d^3)$.

Problema *NP*-completo.

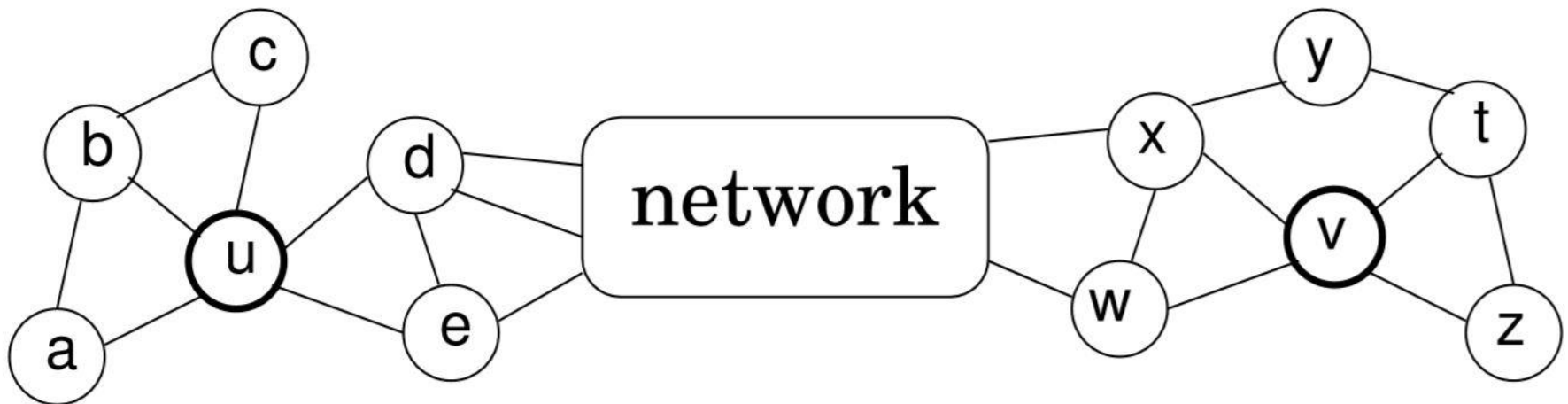


Approcci che usano il ML: DeepWalk, node2vec e **struc2vec**

La *structural similarity* corrisponde ad un concetto di *simmetria* nel quale i nodi di una rete vengono identificati in base alla struttura della rete stessa e tramite relazioni con altri nodi.

struc2vec

- 1) è un framework flessibile
- 2) apprende le *latent representations* per la somiglianza strutturale di ogni nodo
- 3) È molto performante anche in caso di forte rumore (archi eliminati). $O(k^*|V|^3)$



- 1) Determina la *somiglianza strutturale* tra ogni coppia di nodi per dimensioni di neighborhood crescenti.

$$f_k(u, v) = f_{k-1}(u, v) + g(s(R_k(u)), s(R_k(v)))$$

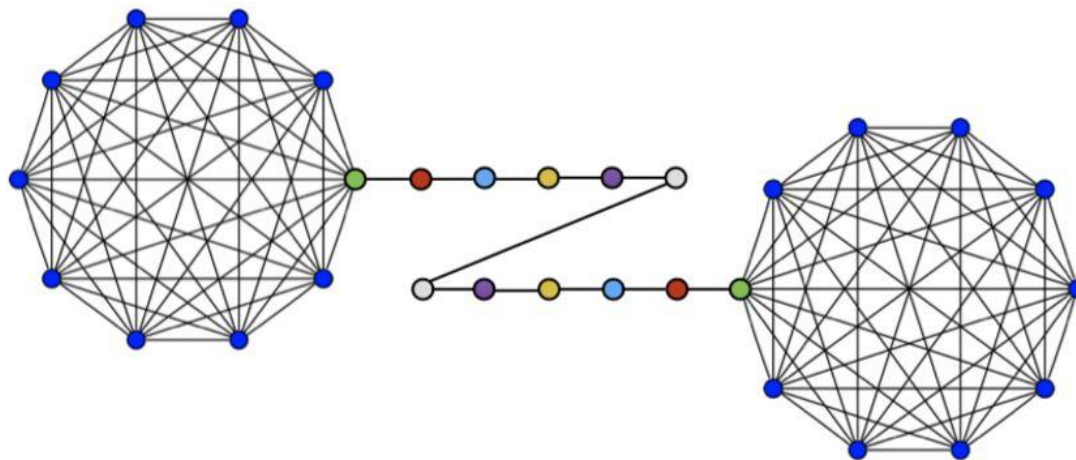
- 2) Costruisce un grafo multilivello pesato. Ogni livello è un grafo completo composto da tutti i nodi del grafo originale, il peso tra ogni coppia di nodi è $w_k(u, v) = e^{-f_k(u, v)}$

Ogni nodo è collegato col corrispondente nel layer inferiore e superiore con arco di peso $w(u_k, u_{k+1}) = \log(\Gamma_k(u) + e)$

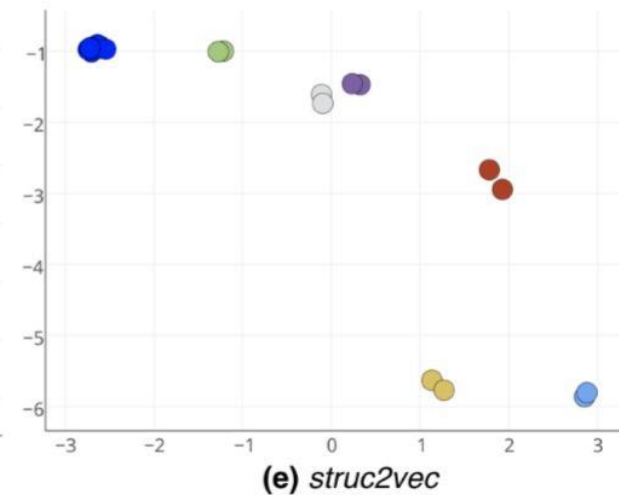
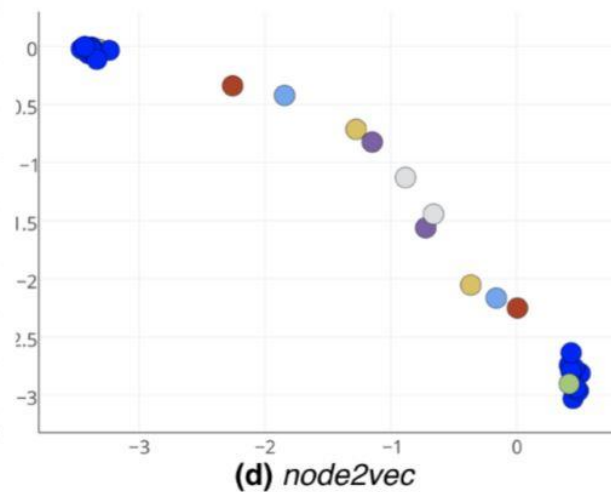
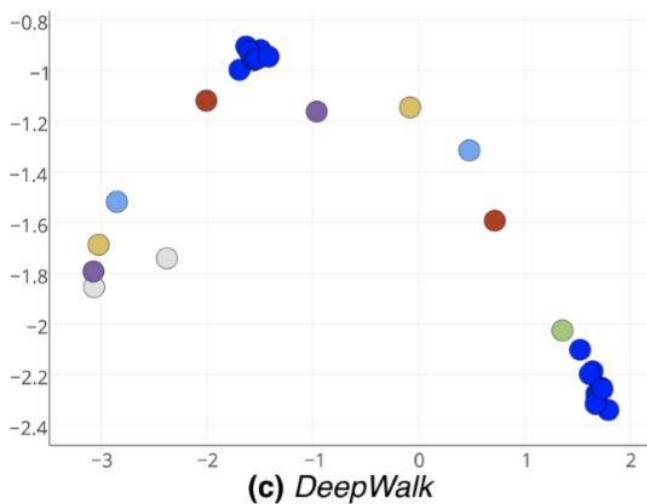
- 3) Si attraversa più volte il grafo multilivello con un algoritmo semi-randomico, generando diverse sequenze di nodi.
- 4) Da queste sequenze di nodi si utilizza una tecnica di *unsupervised learning* per generare la *latent representations* (un vettore) di ogni nodo.

2 nodi strutturalmente simili hanno *latent representations* vicine.

struc2vec vs *DeepWalk* e *node2vec*

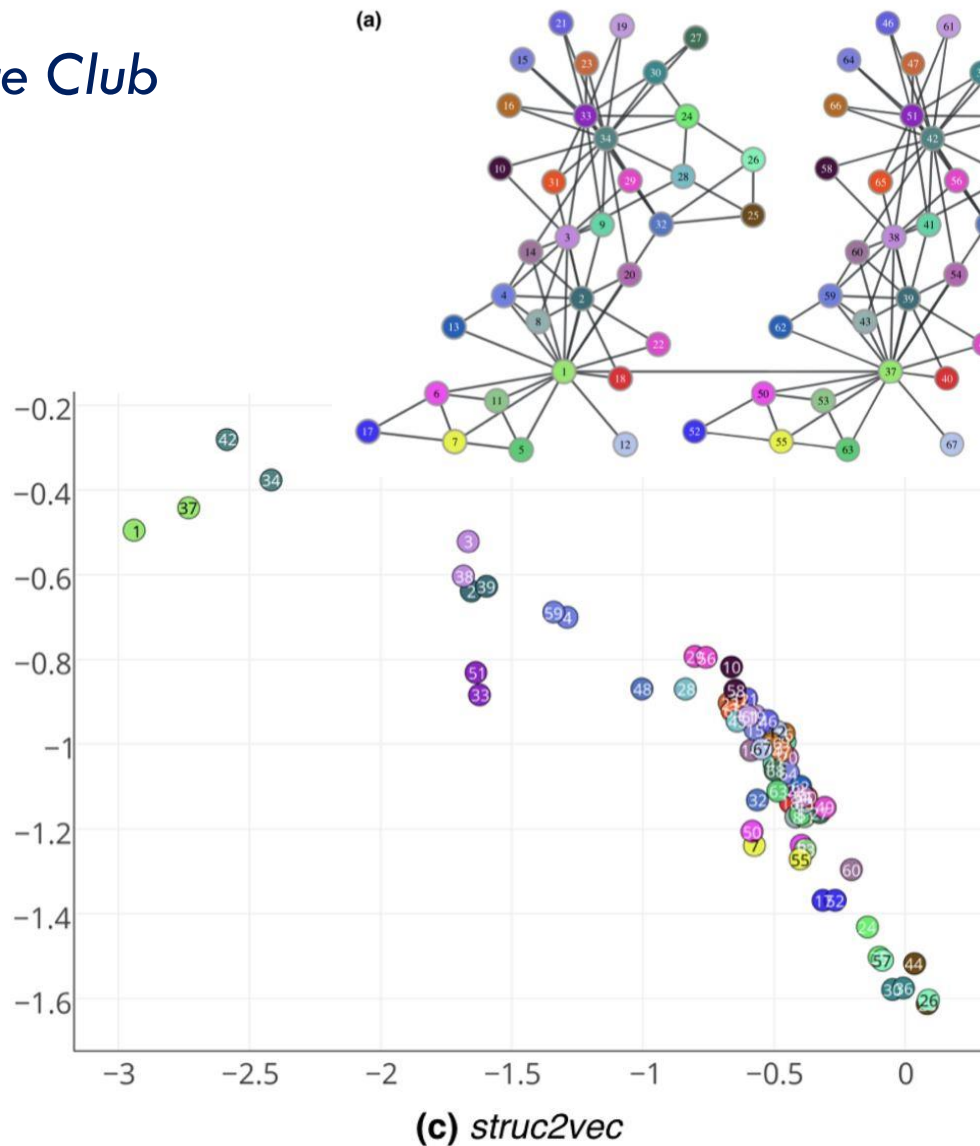
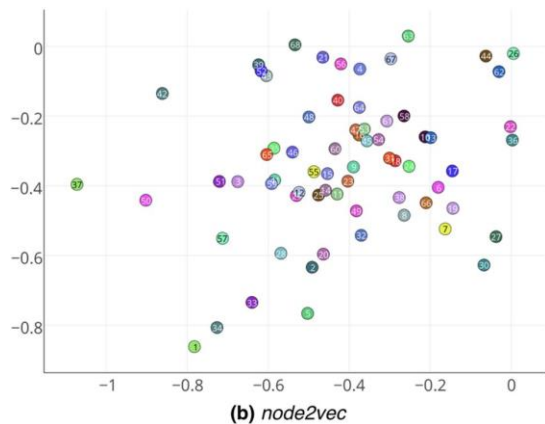
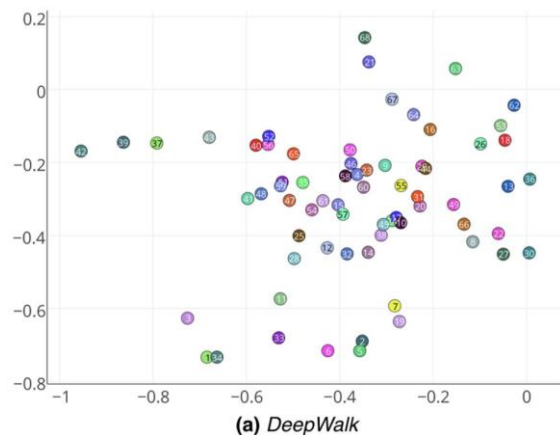


Barbell graph



struc2vec vs *DeepWalk* e *node2vec*

Zachary's Karate Club



Network Alignment: allineare/trovare similitudini tra grafi.

In particolare tra le **Reti di Interazione Proteina-Proteina**

Il corpus di dati PPI è cresciuto esponenzialmente.

4 metodi:

- **MTGO:** 13 citazioni
- **Isorank:** 500 citazioni ma in calo dal 2008
- **Lgraal:** empiricamente migliore di Isorank $O(|V|^3 + |V|^2 \times d^3)$
- **Struc2vec:** complesso ma le sue componenti possono essere modificate, $O(k^*|V|^3)$, sempre più citato (>300 dal 2017)

Futuro: ML (struc2vec) per via della NP-Completezza