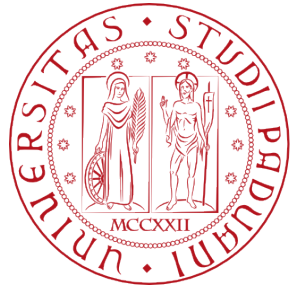
Conjunctive Disjunctive Node Kernel

Dinh Tran-Van[1], Alessandro Sperduti[1] and Fabrizio Costa[2]

1- Department of Mathematics, University of Padova, Italy
2- Department of Computer Science, University of Exeter, UK

April 27^{th}, 2017



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Outline

Motivation

How to improve the performance of disease-gene association predictive systems?

- Disease-gene association predictive systems are often based on the notion of relation between genes.
- A common strategy is to encode gene-gene relations as a graphs and employ graph-based techniques to make inferences.
- A key to determine the system's performance is the similarity measurement

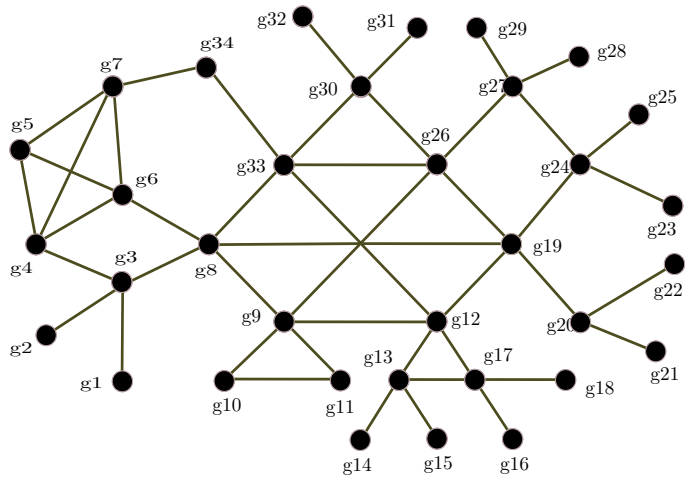


Figure: Genetic graph

How to improve the performance of disease-gene association predictive systems?

- Graph node kernels are normally used to measure node similarities.
- Most node kernels are based on transitive properties and have limitations:
  - low discriminative capacity
  - prefering dense graphs, they show poor performances in case of sparse graphs.

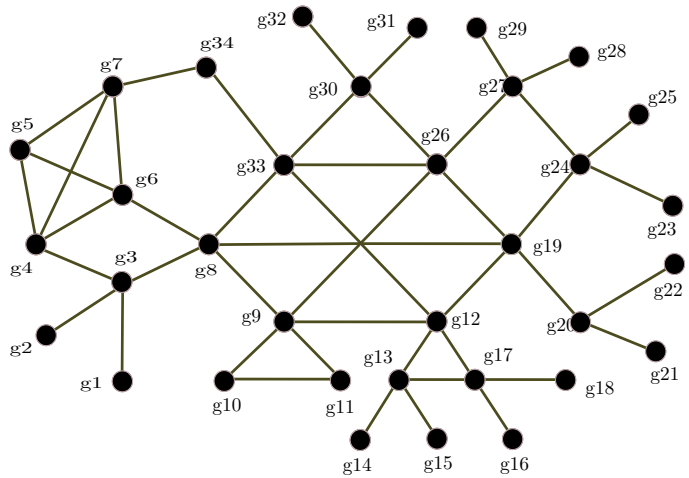


Figure: Genetic graph

Method

Proposed kernel

We propose Conjunctive Disjunctive Node Kernel (CDNK) which is an instance of decompositional graph kernel (DGK) [1] and a modification of NSPDK kernel [2].

Advantages

- It takes advantage from NSPDK kernel which can explicitly model the configuration of nodes' context.
- It contains a decompsition procedure which transforms graph into a collection of linked sparse subgraphs so that DGK can efficiently work. Therefore, we can use take dense or sparse graphs as the input of our kernel.

Notations

- $G = (V, E)$: an undirected, labeled graph with node set $V(E)$, and edge set $E(G)$
- $\mathcal{D}(u, v)$: shorest distance between u and v
- $N_r(v) = \{u | \mathcal{D}(v, u) \leq r\}$: neighborhood set with radius r
- $\mathcal{N}_r^v$: subgraph formed by nodes and edges with endpoints in $N_r(v)$

Neighborhood Subgraph Pairwise Distance Kernel (NSPDK)

- NSPDK is an instance of decompositional kernels
- $R_{r,d}(A_u, B_v, G)$ is true if $A_u \cong N_r^u$, $B_v \cong N_r^v$ and $\mathcal{D}(u,v) = d$
- $R_{r,d}^{-1}(A_u, B_v, G) = \{A_u, B_v | R_{r,d}(A_u, B_v, G) = true\}$
- $\kappa_{r,d}(G, G') = \sum_{\substack{A_u, B_v \in R_{r,d}^{-1}(G) \\ A'_{u'}, B'_{v'} \in R_{r,d}^{-1}(G')}} \mathbf{1}_{A_u \cong A'_{u'}} \cdot \mathbf{1}_{B_v \cong B'_{v'}}$, where $\mathbf{1}_{A \cong B}$ is the *exact matching function* that returns 1 if A is isomorphic to B and 0 otherwise.
- $K(G, G') = \sum_r \sum_d \kappa_{r,d}(G, G')$
  for efficiency reasons, the values of r and d are upper bounded to a given r^* and d^*, respectively.

Node Labeling

We propose to use discretized functional annotation based on gene ontology.

- Each gene is representated as a vector of go-terms
- Cluster genes into a given number of clusters
- Genes are labeled as their cluster class identifiers

Method

Node Labeling

We propose to use discretized functional annotation based on gene ontology.

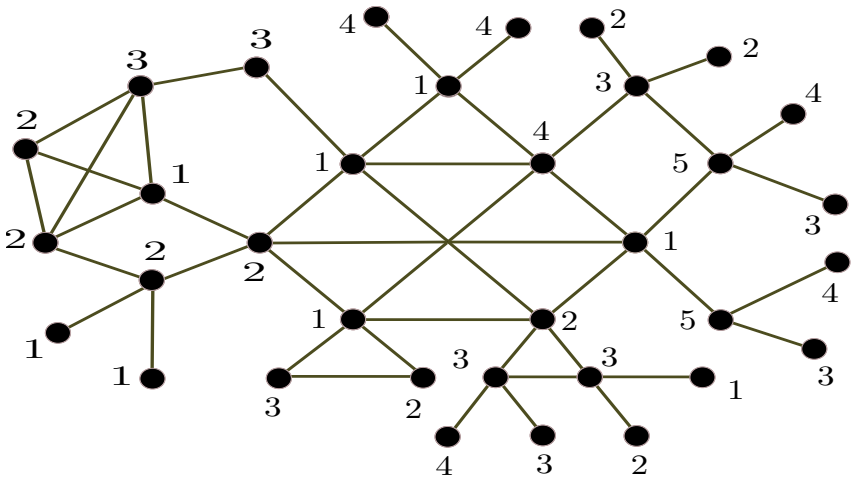
- Each gene is representated as a vector of go-terms
- Cluster genes into a given number of clusters
- Genes are labeled as their cluster class identifiers

Graph Decomposition

Transforming graph in a collection of linked sparse sub-graphs in which we use two types of links: *conjunctive* and *disjunctive*.

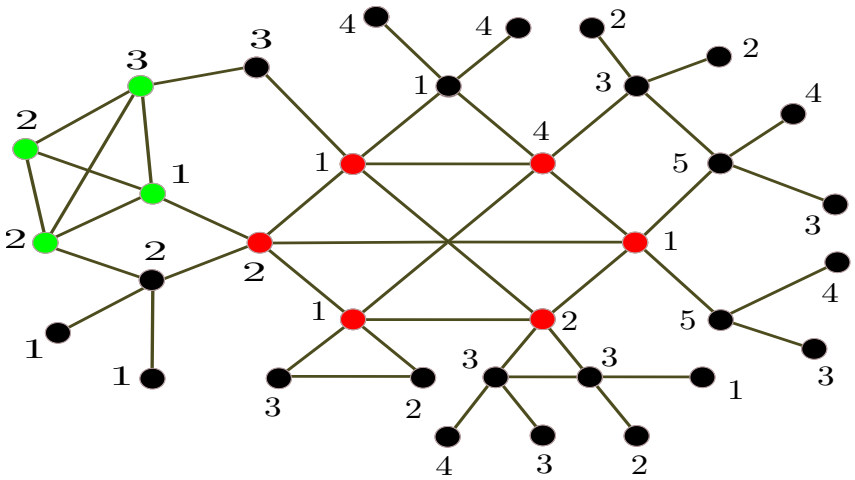
- *conjunctive:* used for considering distance between nodes
- *disjunctive:* used for connecting sparse subgraphs

Method



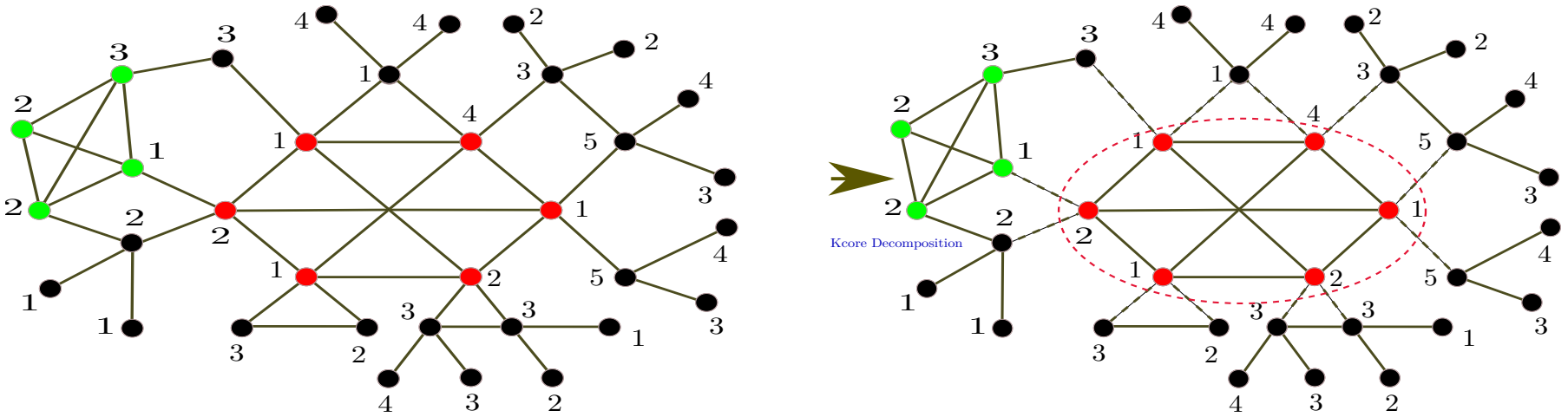
- **Iterative Kcore**: form a collection of linked subgraphs
- **Clique Decomposition**: similarly treat nodes belong same clique

Method



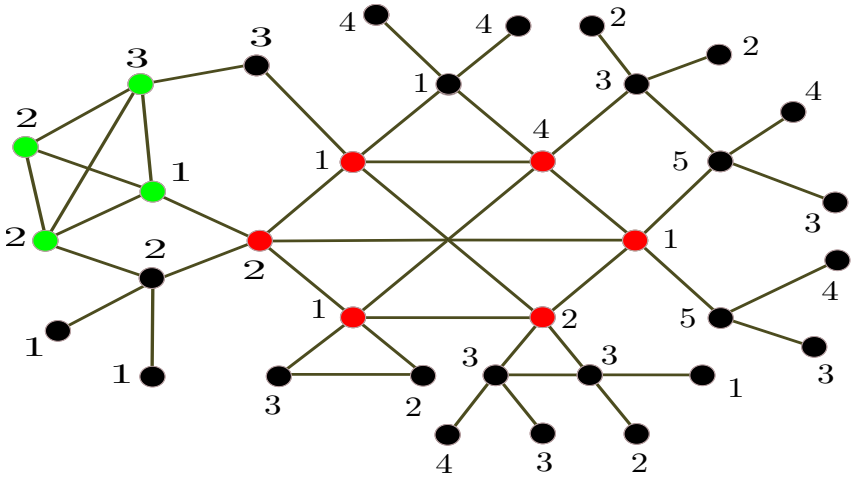
- **Iterative Kcore**: form a collection of linked subgraphs
- **Clique Decomposition**: similarly treat nodes belong same clique

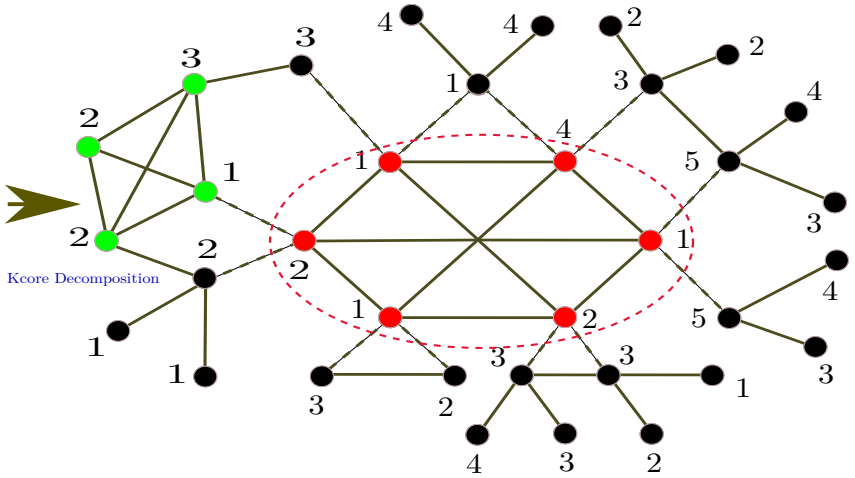
Method



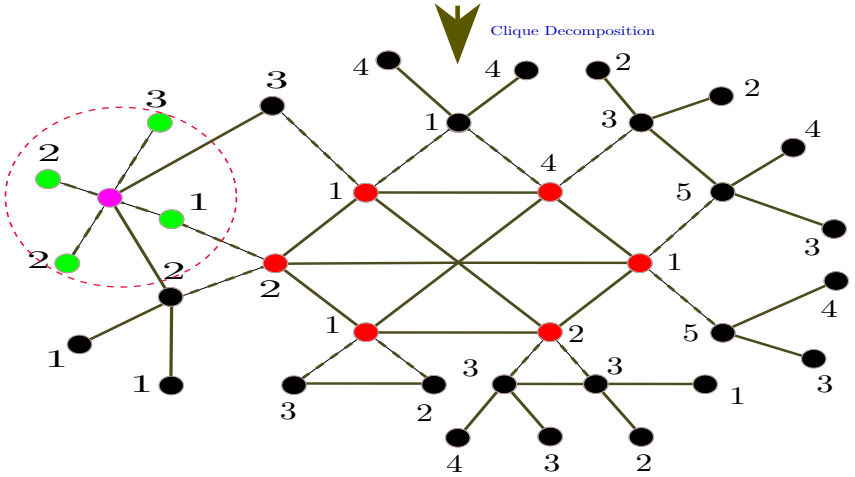
- **Iterative Kcore**: form a collection of linked subgraphs
- **Clique Decomposition**: similarly treat nodes belong same clique

Method





- **Iterative Kcore**: form a collection of linked subgraphs
- **Clique Decomposition**: similarly treat nodes belong same clique



Method

We first define:

- Conjunctive relation: $R^{\wedge}_{r,d}(A_u, B_v, G_u)$ is true if $A_u \cong N^u_r$, $B_v \cong N^v_r$ and $\mathcal{D}(u, v) = d$
- Disjunctive relation: $R^{\vee}_{r,d}(A_u, B_v, G_u)$ is true if $A_u \cong N^u_r$, $B_v \cong N^v_r$ and $\mathcal{D}(w, v) = d$, (u, w) is a disjunctive edge.
- $\kappa_{r,d}(G_u, G_{u'}) =$
$$\sum_{\substack{A_u, B_v \in {R^{\wedge}_{r,d}}^{-1}(G_u) \\ A'_{u'}, {B'}_{v'} \in {R^{\wedge}_{r,d}}^{-1}(G_{u'})}} \mathbf{1}_{A_u \cong A'_{u'}} \cdot \mathbf{1}_{B_v \cong B'_{v'}} + \sum_{\substack{A_u, B_v \in {R^{\vee}_{r,d}}^{-1}(G_u) \\ A'_{u'}, {B'}_{v'} \in {R^{\vee}_{r,d}}^{-1}(G_{u'})}} \mathbf{1}_{A_u \cong A'_{u'}} \cdot \mathbf{1}_{B_v \cong B'_{v'}}.$$

CDNK is defined as: $K(G_u, G_v) = \sum_r \sum_d \kappa_{r,d}(G_u, G_v)$.

Evaluation

We evaluate performance of kernels by employing gene prioritization on 12 diseases and using two datasets (followed [3]).

- **Kernels**: K1: Diffusion kernel [4], K2: Markov difussion kerel [5], K3: Markov exponential diffusion kernel [3], K4: Regularized Laplacian kernel [6], K5: CDNK.
- **Datasets**: *BioGPS* - a gene co-expression network (7311 nodes, 911,294 edges), and *Pathways* - encode gene common pathway relations (7311 nodes, 2,254,822 edges).

Table 1: Performance of kernels in term of average AUC and order ranking over all diseases

	BioGPS					**Pathways**				
Kenels	K1	K2	K3	K4	K5	K1	K2	K3	K4	K5
$\overline{AUC}$	66.6	63.5	67.8	67.5	73.3	62.7	70.1	75.3	75.5	76.5
$\overline{Rank}$	3.5	4.3	2.8	2.5	2.0	4.8	3.9	2.5	2.0	1.8

Conclusion

We have proposed Conjunctive Disjuctive graph node kernel that:

- efficently exploit graph structure to form high discriminative node similarity measurement
- includes a decomposition procedure which allows it to process with both sparse and dense graphs
- shows state of the art performance

Reference

1. Haussler, D. Convolution kernels on discrete structures. Technical report, Department of Computer Science, University of California at Santa Cruz, 1999.
2. Fabrizio C., et al, Fast neighborhood subgraph pairwise distance kernel. Proceedings of the 26th, International Conference on Machine Learning. Omnipress, 2010.
3. Chen, B., et al, Disease gene identification by using graph kernels and Markov random fields. Science China Life Sciences (2014).
4. Kondor, R., et al, Diffusion kernels on graphs and other discrete input spaces. ICML. Vol. 2. 2002.
5. Fouss F, et al, An experimental investigation of graph kernels on a collaborative recom- mendation task. Proceedings of the 6th international conference on data mining 2006. ICDM 2006.
6. Chebotarev P., et al, The matrix forest therem and measuring relations in small social groups. Automation and Remote Control 1997.

THANKS FOR LISTENING