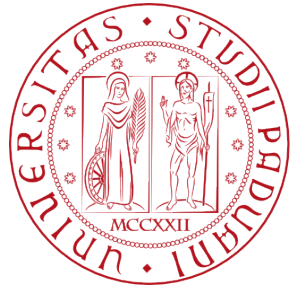
Conjunctive Disjunctive Node Kernel

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

you can delete this slide
as it is not informative

Motivation

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

first define what disease-gene association is

- 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

and use a picture to show how to use info on nearby genes to predict that connected genes are also related

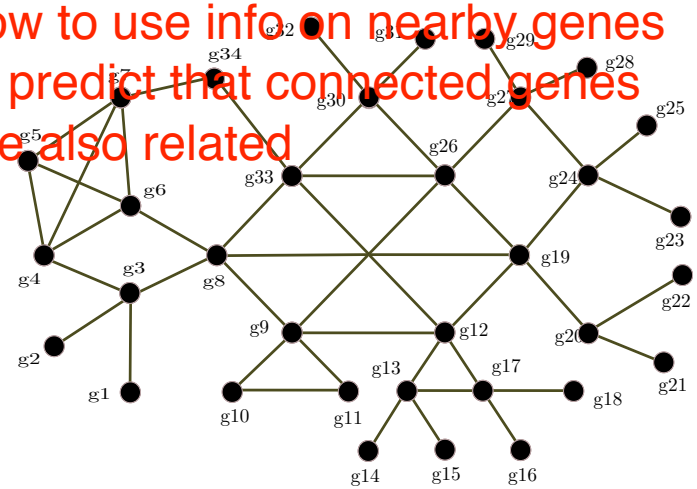


Figure: Genetic graph

in general do not use more than two lines of text per bullet point

Motivation

the title should be used to indicate what is the topic of the slide: here, diffusion kernels

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.

visualize the way diffusion kernels work: use the picture of a gaussian that decreases on distant nodes

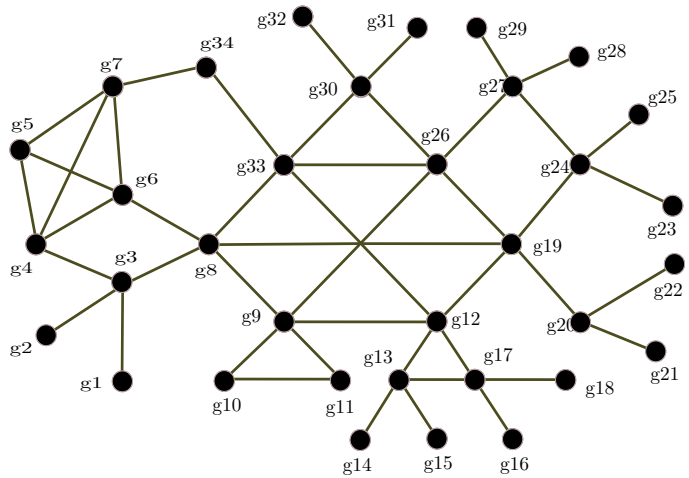


Figure: Genetic graph

this picture is not changing!
every picture should represent
the idea expressed in the current slide

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 each nodes' context.
- ~~It contains a decompsition procedure which~~ transform~~s~~ 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.~~

in the next slides explain all these concepts

Method

do not use formulas in the slides

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 $N_r(v)$

replace this and next slide with:
- description of what is a decompsitional kernel
- description of context and NSPDK features
describe using pictures not formulas!

remove

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.

Method

Node Labeling

We propose to use discretized functional annotation based on ~~gene~~ ontology. Gene Ontology (GO)

- Each gene is representated as a ~~vector~~ bag of ~~go~~ GO-terms
- Cluster genes in~~to a given number of~~ k clusters
- Gene~~s are~~ labele~~d as~~ is their cluster ~~class~~ identifier~~s~~

first explain *why* we need node labels
to characterize the context (otherwise the features are considering only the relational structure and not the association to certain functions)
then say *how* we do it

Method

do not repeat information from previous slides

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

say why we do the decomposition step:
i.e. to get sparse graphs
say what a sparse graph is

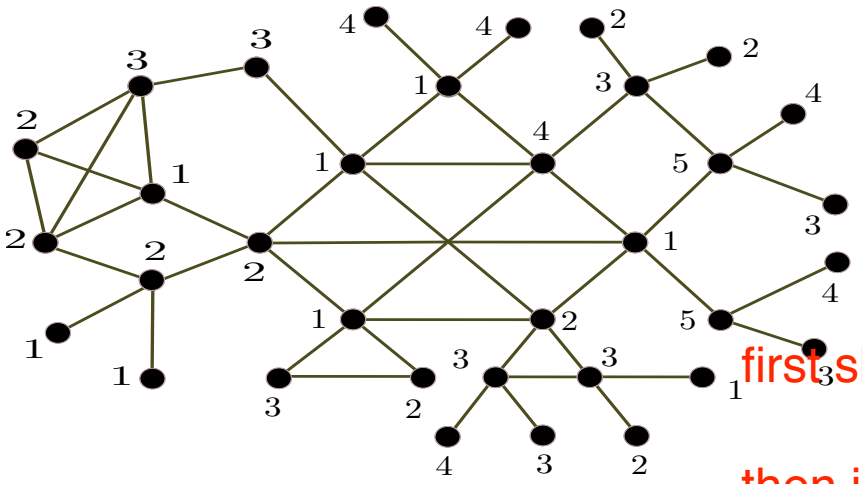
Graph Decomposition

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

dont introduce conj disj now, it's too early

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

Method



first show the notion of k-Core alone with 2 pictures

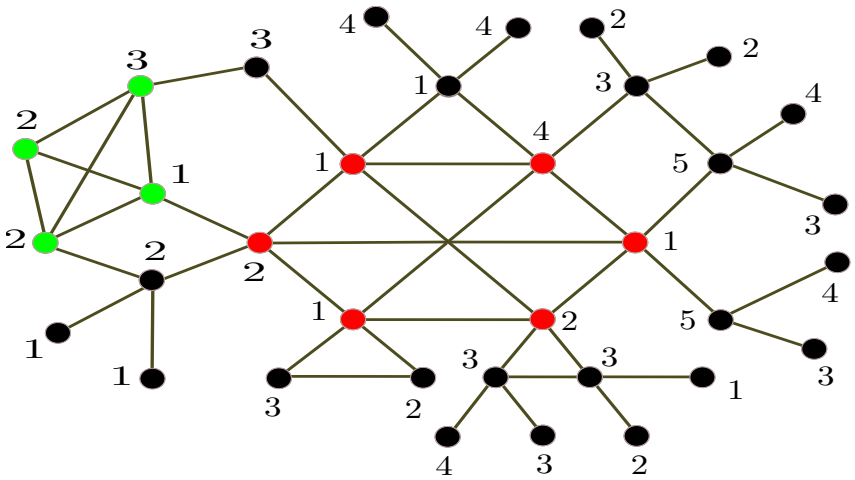
then in separate slides
show the concept of clique decomposition

- **Iterative Kcore**: form a collection of linked subgraphs

at the beginning define k-Core and clique in words

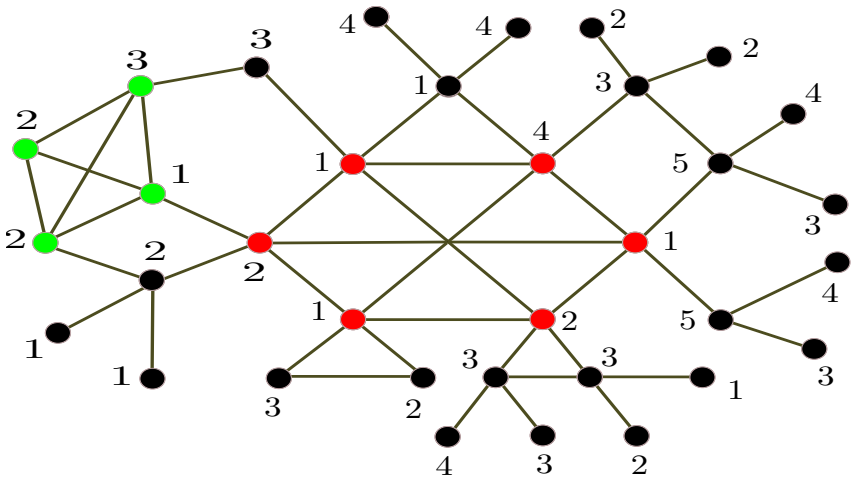
- **Clique Decomposition**: similarly treat nodes belong same clique

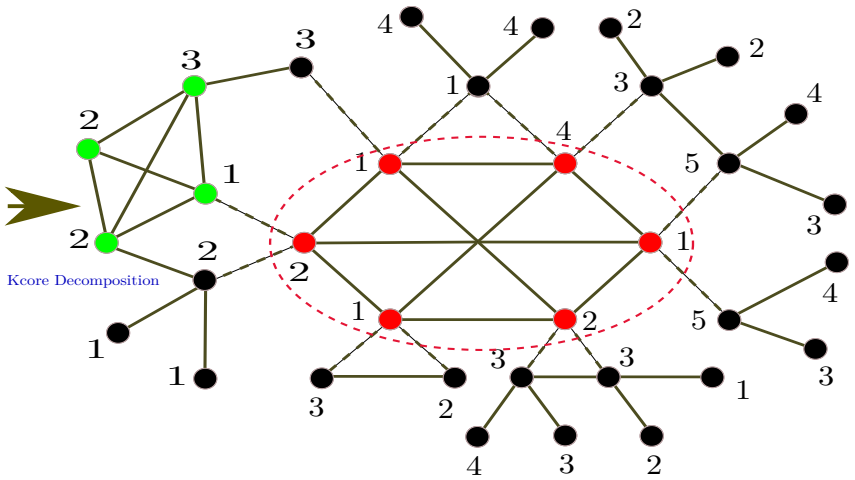
Method



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

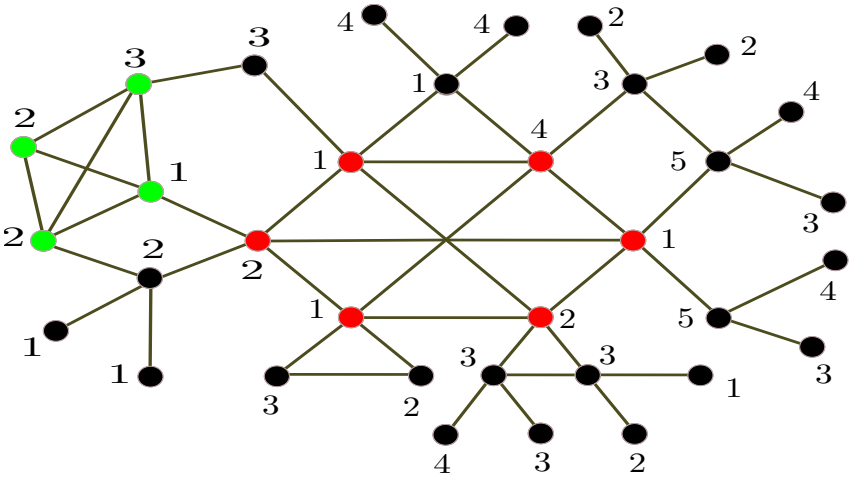
Method

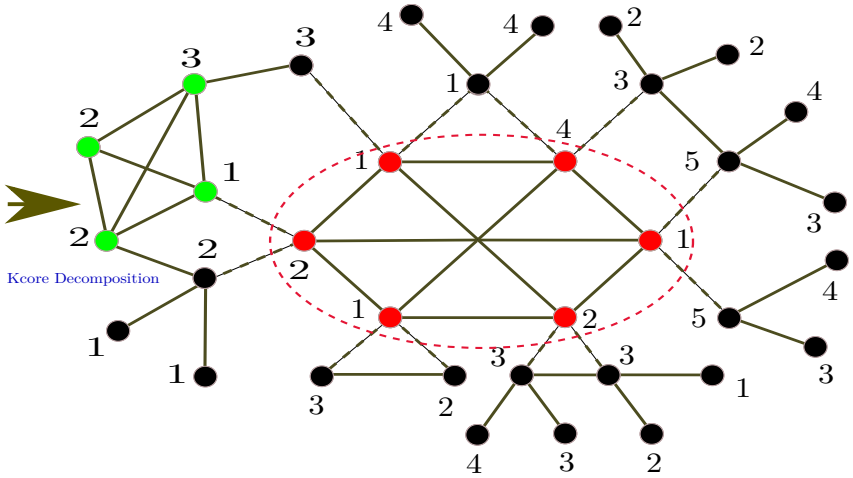




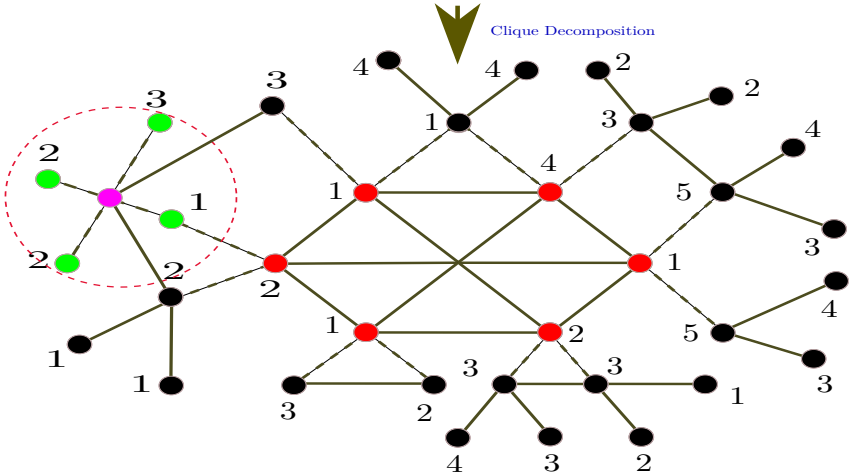
- **Iterative Kcore**: form a collection of linked subgraphs
- **Clique Decomposition**: similarly treat nodes belong same clique

Method





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



Method

replace this with a picture that shows an example of feature
i.e. pairs or neighborhood graphs
first extracted on the original dense network
and then using the same roots but on the decomposed network

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]).

make a vertical list of disease names

- **Kernels**: K1: Diffusion kernel [4], K2: Markov difussion kerel [5], K3: Markov exponential diffusion kernel [3], K4: Regularized Laplacian kernel [6], K5: CDNK. write the kernels as a vertical list
- **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

put table on new slide
add description of predictive task in ML terms

	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