Disease Module Identification and Analysis

Robert Bosch Centre for
Data Science and Artificial Intelligence
(RBC-DSAI)

Beethika Tripathi [1,3,4,5]   Himanshu Sinha [2,3,4]   Karthik Raman [2,3,4,5]   Balaraman Ravindran [1,3,4,5]

[1] Department of Computer Science and Engineering, Indian Institute of Technology (IIT) Madras, Chennai, INDIA
[2] Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, IIT Madras
[3] Robert Bosch Centre for Data Science and Artificial Intelligence(RBC-DSAI), IIT Madras
[4] Initiative for Biological Systems Engineering, IIT Madras
[5] Interdisciplinary Laboratory for Data Sciences, IIT Madras

Motivation

- Physiological and disease processes are driven by multiple genes that interact within molecular modules or pathways.
- Identification of these modules, is a prerequisite for understanding disease-gene relationship or for a targeted search of drug targets.
- Many heuristics are available to identify such modules but the basic underlying connectivity patterns remain largely unexplored.

Introduction

The idea is genes that are co-localized in the interaction networks are more likely to be functionally correlated. ss

- Disease Modules are typically small (3-100) in comparison to the size of the networks, so off-the-shelf algorithms fail.
- Inherent noise present in network due to their nature of curation, so important to efficiently integrate heterogeneous sources of information.
- Creating and analysing ground truth disease modules.
- Co-morbidity analysis to find association between different diseases.

Community structure refers to the occurrence of groups of nodes in a network that are more densely connected internally than with the rest of the network. This inhomogeneity of connections suggests that the network has certain natural divisions within it.


Identify modules for each network individually   Identify modules across multiple networks
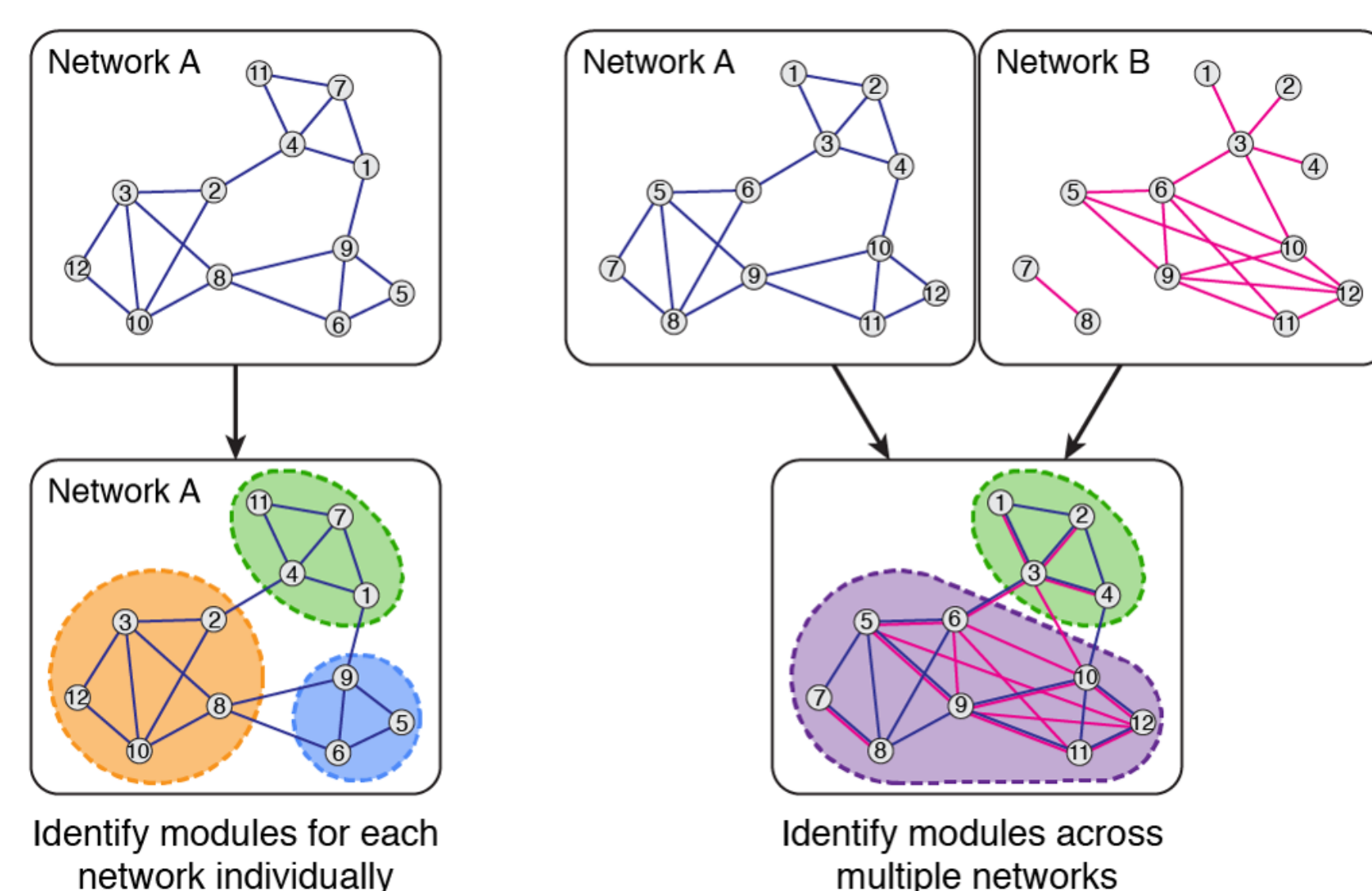
Community Detection

- Determine the scoring function which satisfies the structure of the community that has to be detected.
- **Modularity score**[2] compares the number of edges inside a cluster with the expected number of edges in a random network with the same number of nodes and each node keeps its degree.
- Optimize the score by starting from a single node as community then growing it by adding nodes which maximizes the score.
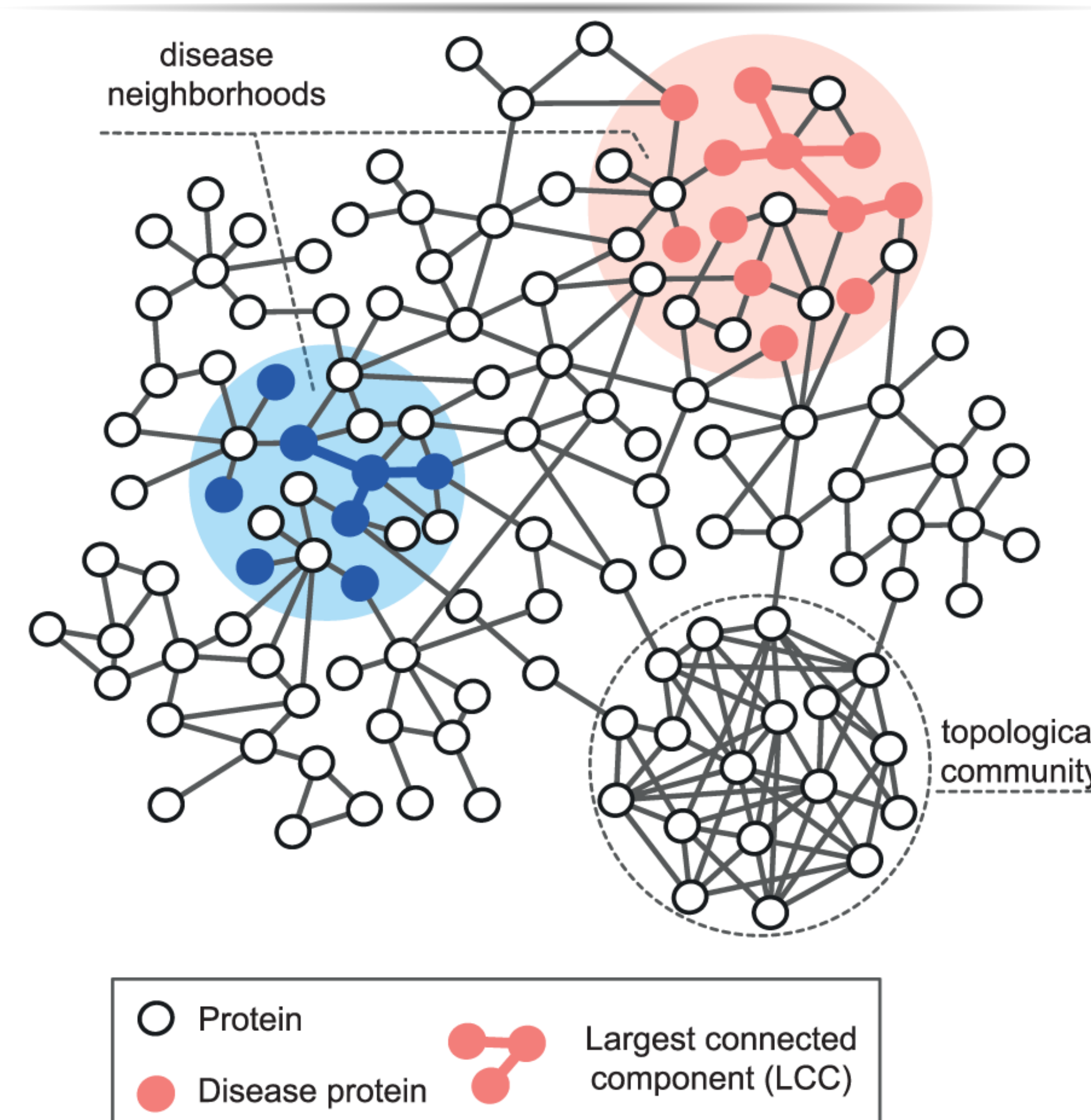
$$ModularityScore(c) = \frac{1}{2m}\Sigma_{i,j\ i\neq j}\left(X_{i,j} - \frac{k_ik_j}{2m}\right)\delta_{c_ic_j}$$

Core Module Identification
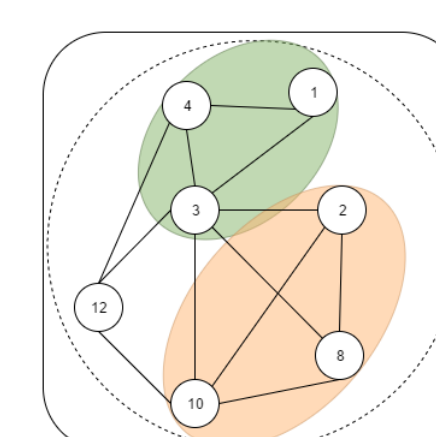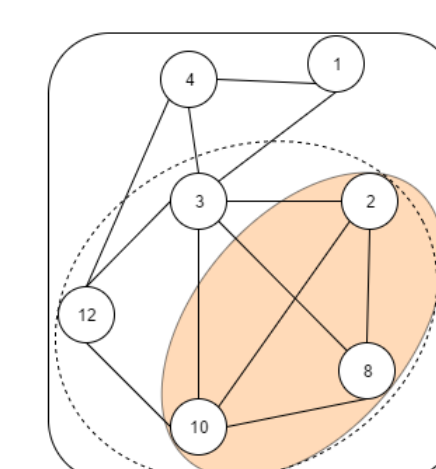
- Perturbing the network & identifying most undisturbed community across all.
- Ensemble of models.
- Core with minimum outgoing edges.
- Multiple core by breaking the larger modules into smaller.

Problems with Community Detection

Creating Gold Standard Disease Module

The algorithm is based on systematic analysis of the network properties of known disease proteins across 180 disorders which is collected from GWAS.

- **Seed Node Selection**: Genes which are significantly associated with a particular trait are selected as disease seed nodes.
- **Seed Node Expansion**[1]: Community starts growing form these seeds by adding nodes from network that have high page rank score associated with given seed node.
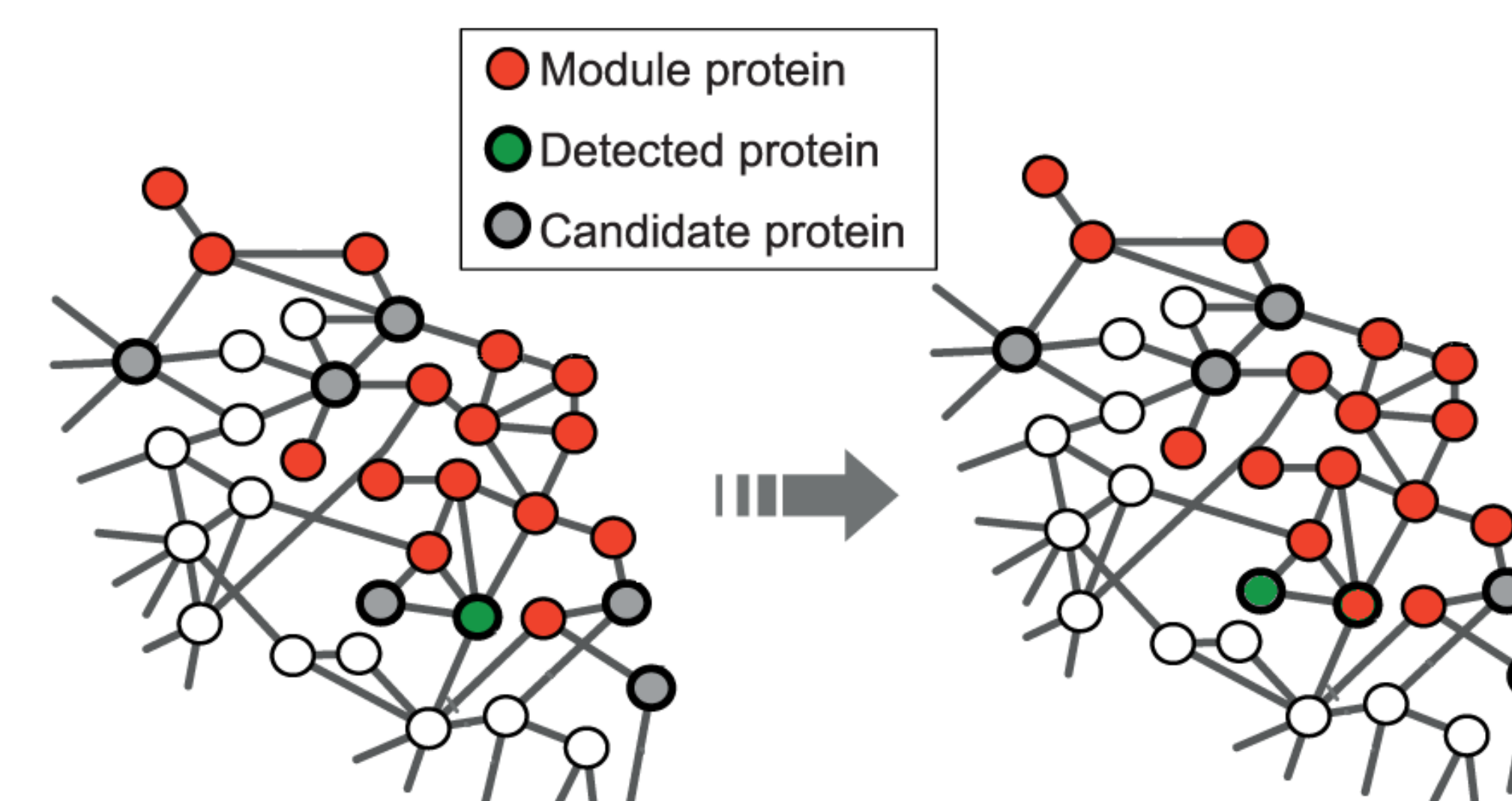
## Information in the Overlap

- Analysing overlaps as communities yields decent enrichment with disease genes.
- Overlaps are most important part of the module and the disease nodes lie mostly on the fringes of the module.
- **Biological justification:** the pleiotropic genes, which have multiple functionalities, should be part of multiple disease modules.

Comorbidity Analysis

Comorbidity is the presence of one or more additional diseases or disorders co-occurring with a primary disorder.

- Comorbidity occurs due to a high overlap of genes between the disease module of different disorders.
- Modules that are getting enriched with multiple diseases used to find out association between the diseases.

Results
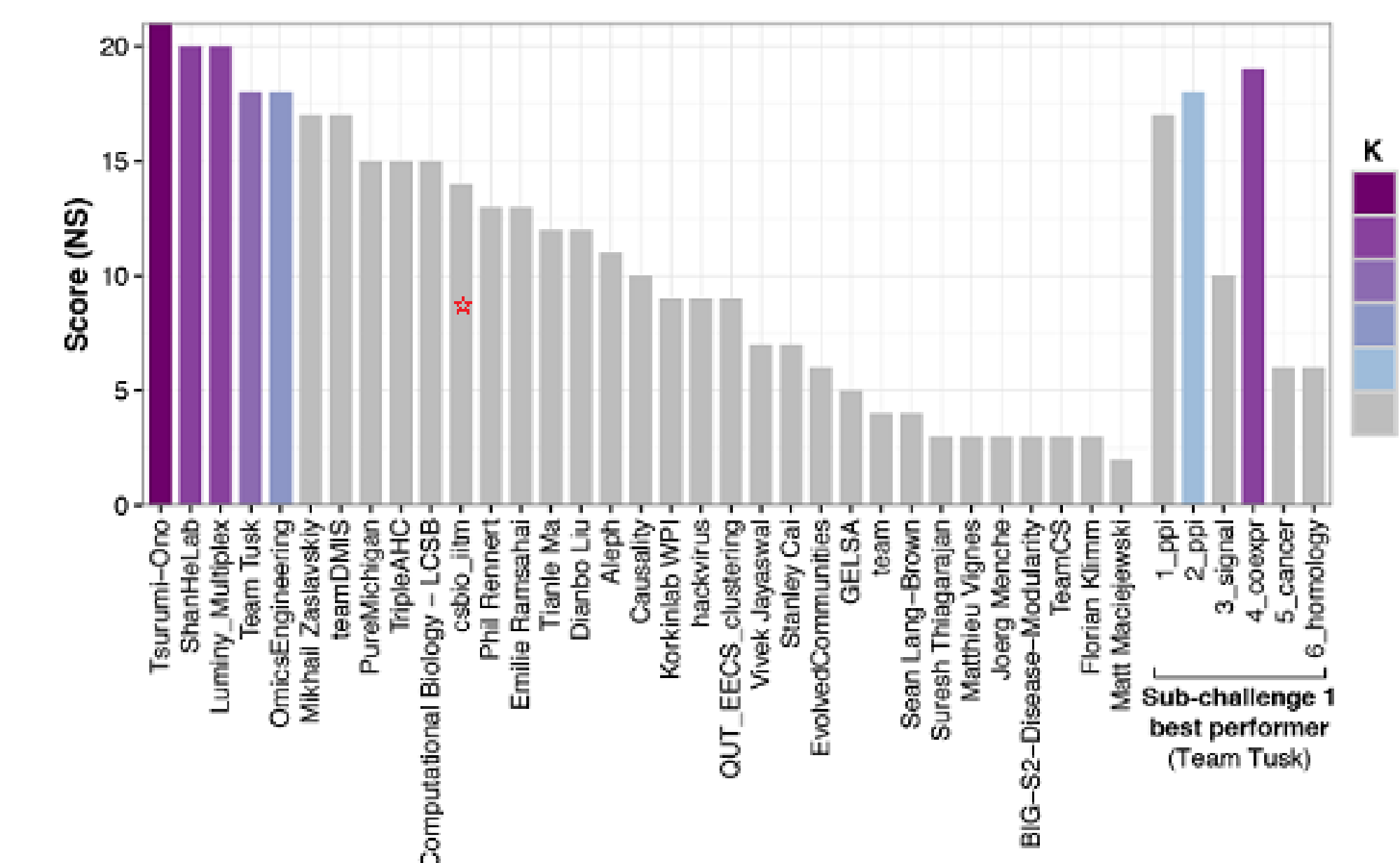
Ranking in Disease Module Identification in DREAM Challenge

Conclusion

- We find that re-clustering large modules, using several heuristics, detects better disease modules.
- Diseases cannot be associated just with topologically dense network communities, instead we identified the interaction significance as the key quantity to characterize the connection patterns among disease proteins.
- Disease genes lie mostly in the fringes of the module rather than to the core of it
- Co-morbid diseases can be identified with disease module sharing same group of genes.

On-going Work

- Incorporating biological knowledge in form of domain information of protein and using information from heterogeneous sources to better understand the disease modules.

References

[1] Taher H Haveliwala.
Topic-sensitive pagerank.
In *Proceedings of the 11th international conference on World Wide Web,* pages 517–526. ACM, 2002.

[2] M E J Newman.
Modularity and community structure in networks.
Proceedings of the National Academy of Sciences of the United States of America, 103(23):8577–8582, 2006.