Network Motifs of Pathogenic Genes in Human Regulatory Network

Michael Colavita
Mentor: Soheil Feizi
Fourth Annual MIT PRIMES Conference
May 18, 2014

Topics

- **Background**
  - Genetics
  - Regulatory Networks
  - The Human Regulatory Network
- Network Motifs
  - Questions and Methods
  - Sparse Disconnect
  - Low Distance Clustering
  - Network Metrics
- Clustering Detection
  - Method
  - Clusters Found

Genetic Background

- Cell's genes have regulatory effects on each other
  - Upregulation
  - Downregulation
- **Transcription factors** control the expression of other genes
- **Target genes** have no regulatory effects
- Both can be subject to regulation by other genes

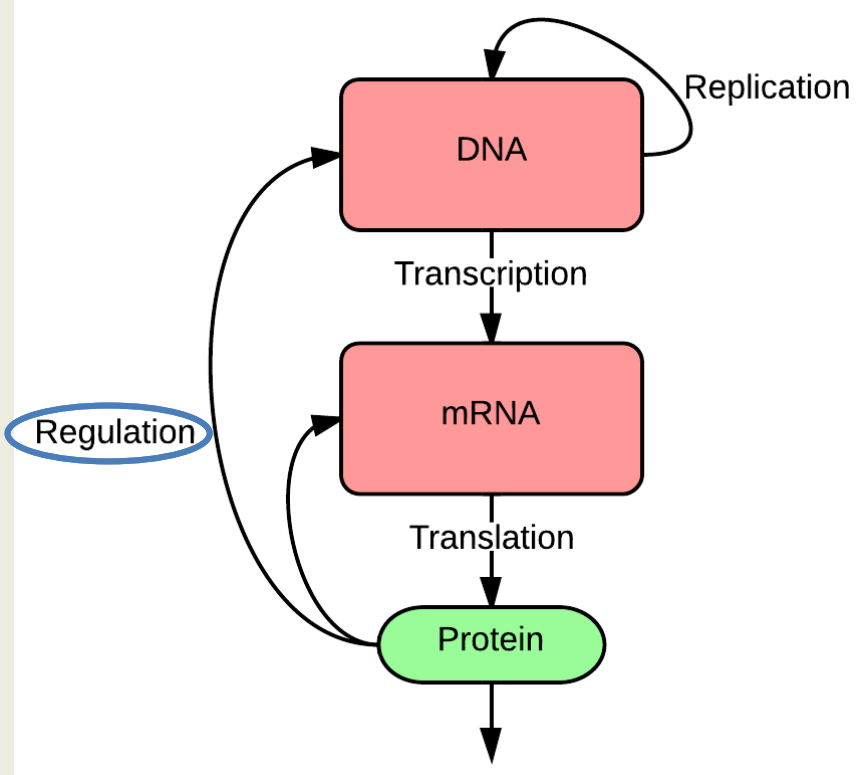


Figure: The central dogma of molecular biology with regulation of gene expression

Genetic Regulatory Networks

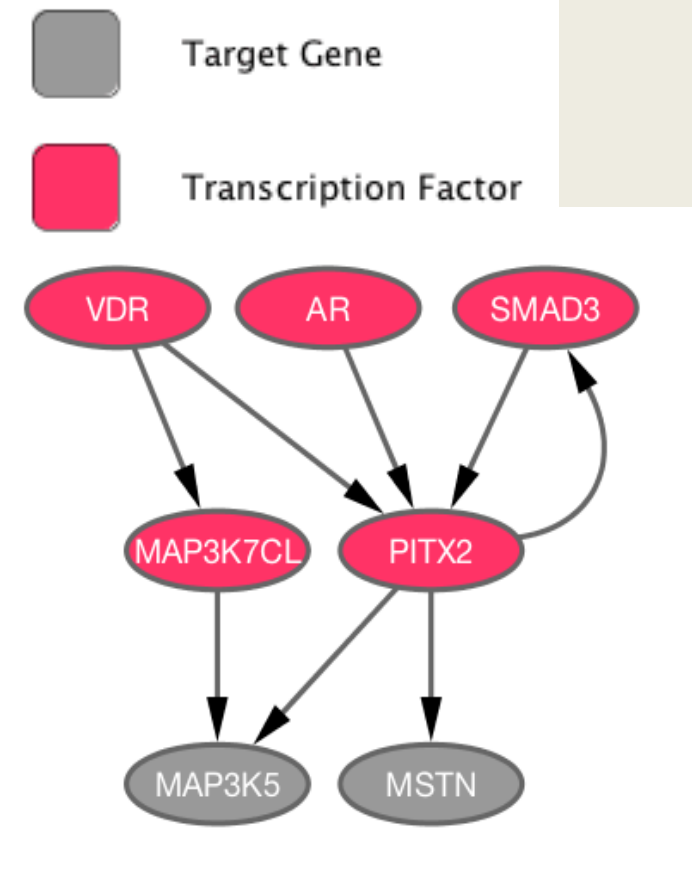


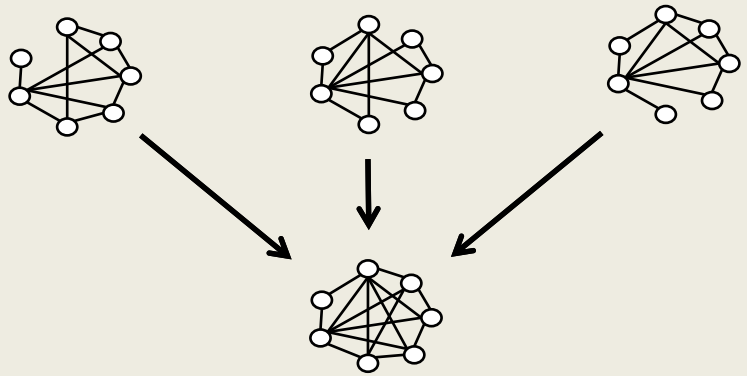
Figure: A sample of the human regulatory network

- Medium for storing regulatory information for computational analysis
- Captures regulatory dynamics of a genome
- Nodes represent genes
- Edges indicate upregulatory effects
  - Edge weights indicate strength of regulatory activity

The Human Regulatory Network

- Primary dataset used for regulation data

- Created by combining datasets into a unified network
  - Co-expression network
  - Motif network
  - ChIP network



- 2757 **transcription factors**
- 16464 **target genes**
- ~1,000,000 regulatory relationships (cutoff = .95)

Topics

- Background
  - Genetics
  - Regulatory Networks
  - The Human Regulatory Network
- **Network Motifs**
  - Questions and Methods
  - Sparse Disconnect
  - Low Distance Clustering
  - Network Metrics
- Clustering Detection
  - Method
  - Clusters Found

Network Motifs of Pathogenic Genes

- Motifs are recurring patterns within the network
  - Patterns in structure
  - Consistent high or low enrichment for given metrics
    - **Indegree**/**Outdegree**
    - **Eigenvector**/**Betweenness Centrality**
    - **Clustering Coefficient**

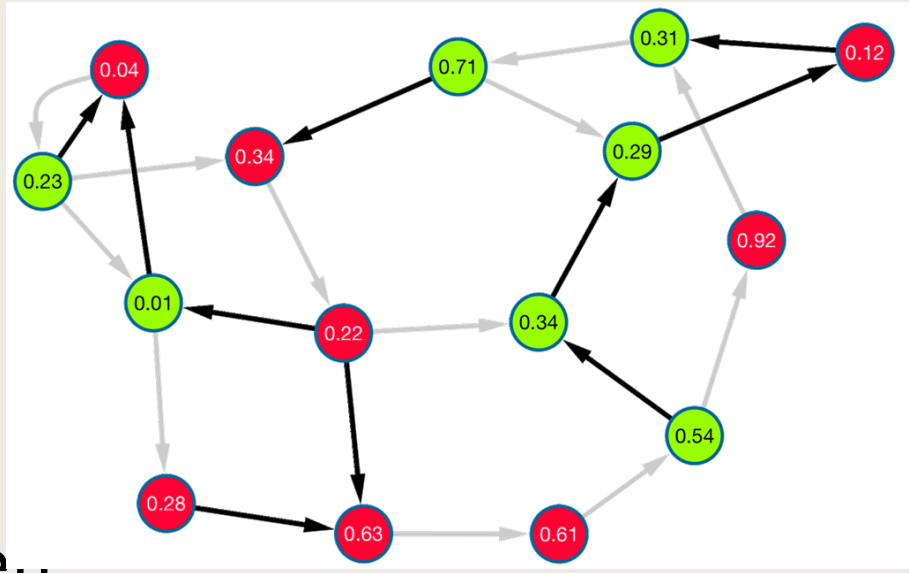
- Do certain network motifs lead to genetic disease through positive feedback?

Motivation for Motif Identification

- Examining motifs of pathogenic genes (dbGaP)
  - Genes associated with genetic disease
- Understanding the regulatory behavior behind genetic diseases
- Investigating larger scale **regulatory structures**
- Possible **regulatory basis** behind genetic disease

Method of Motif Detection

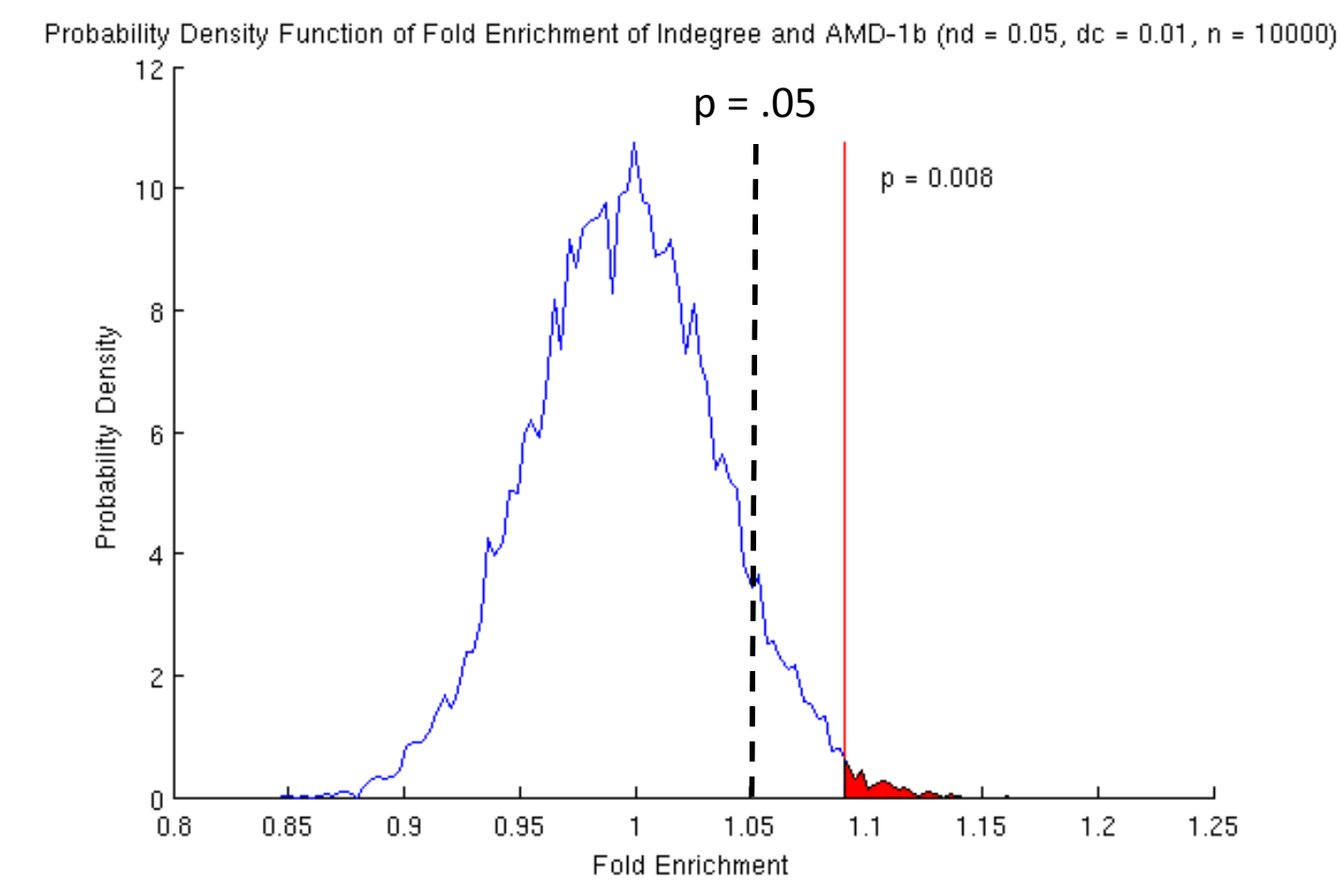
1. Generate a binary network from the top 5% of edges.



2. Compute enrichment of pathogenicity over a given network metric.

3. Compute p-value by comparing this enrichment to that of a randomized disease.

P-value Example

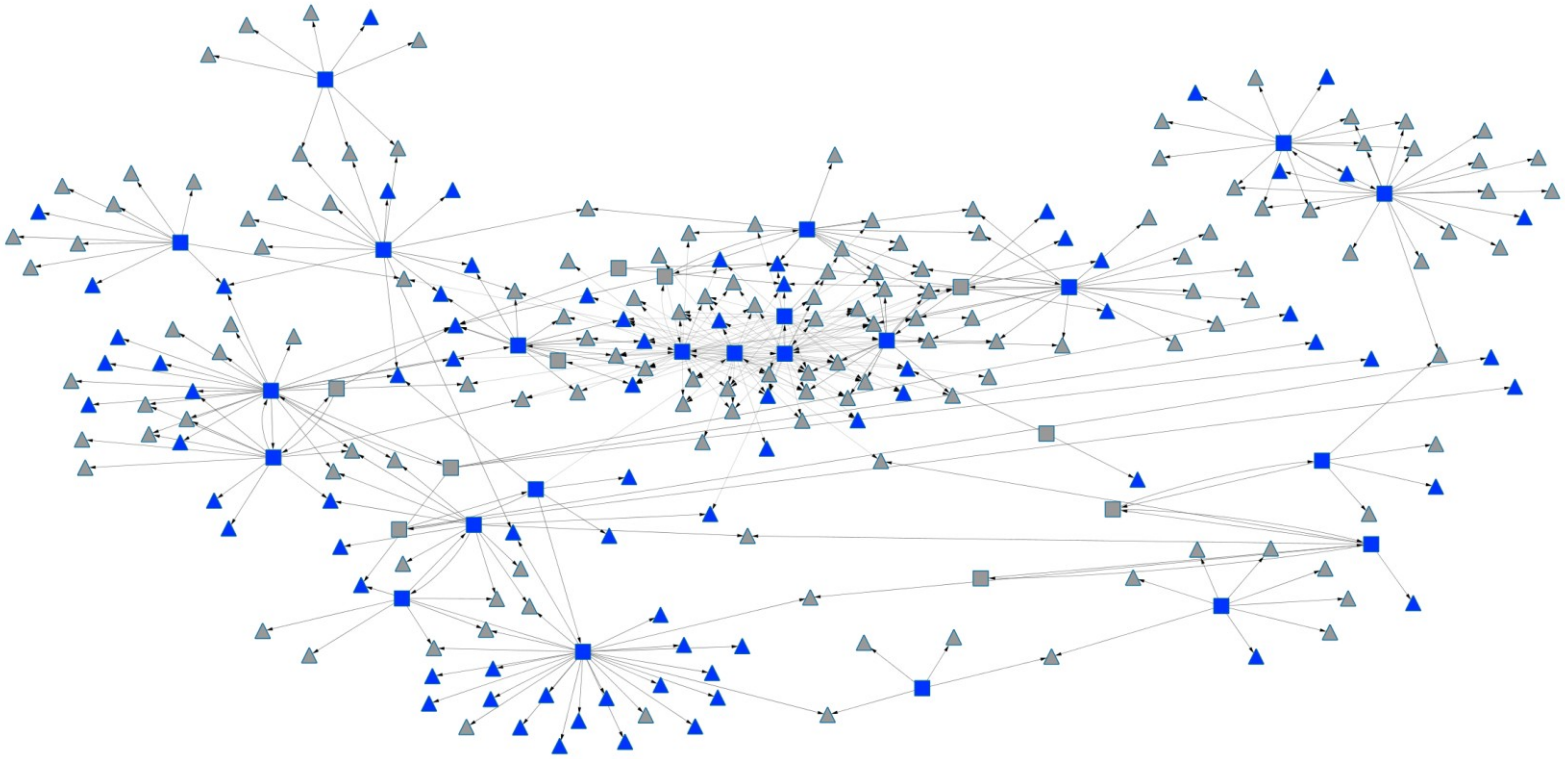


Network Motifs Identified

- Analyzed **45 diseases** in the network of 19,221 genes

- Identified two major motifs so far
  - Sparse disconnect
  - Low distance clustering

Sparse Disconnect Visualization

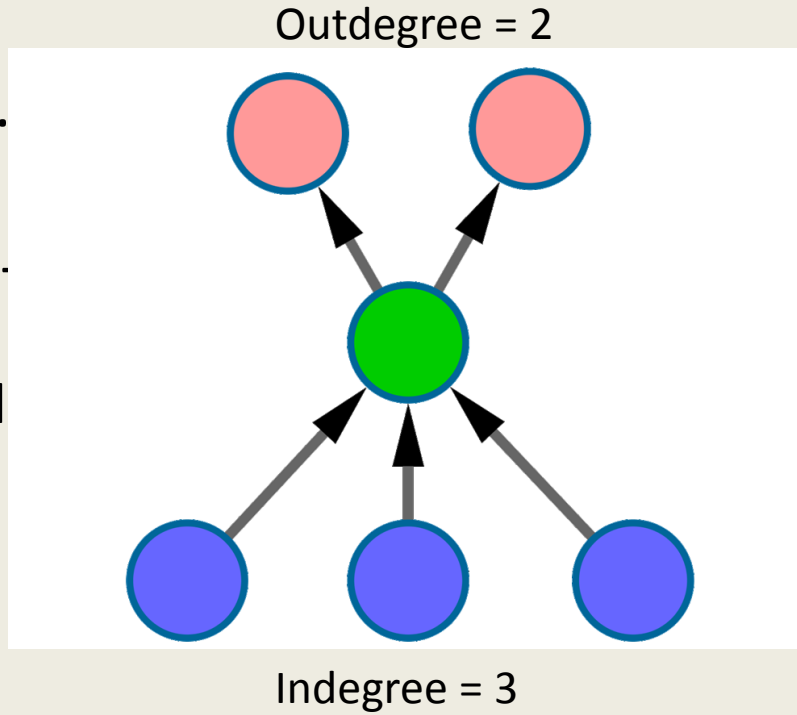


Node Fill Color	Pathogenicity
	Non-pathogenic (p > .01)
	Pathogenic (p < .01)

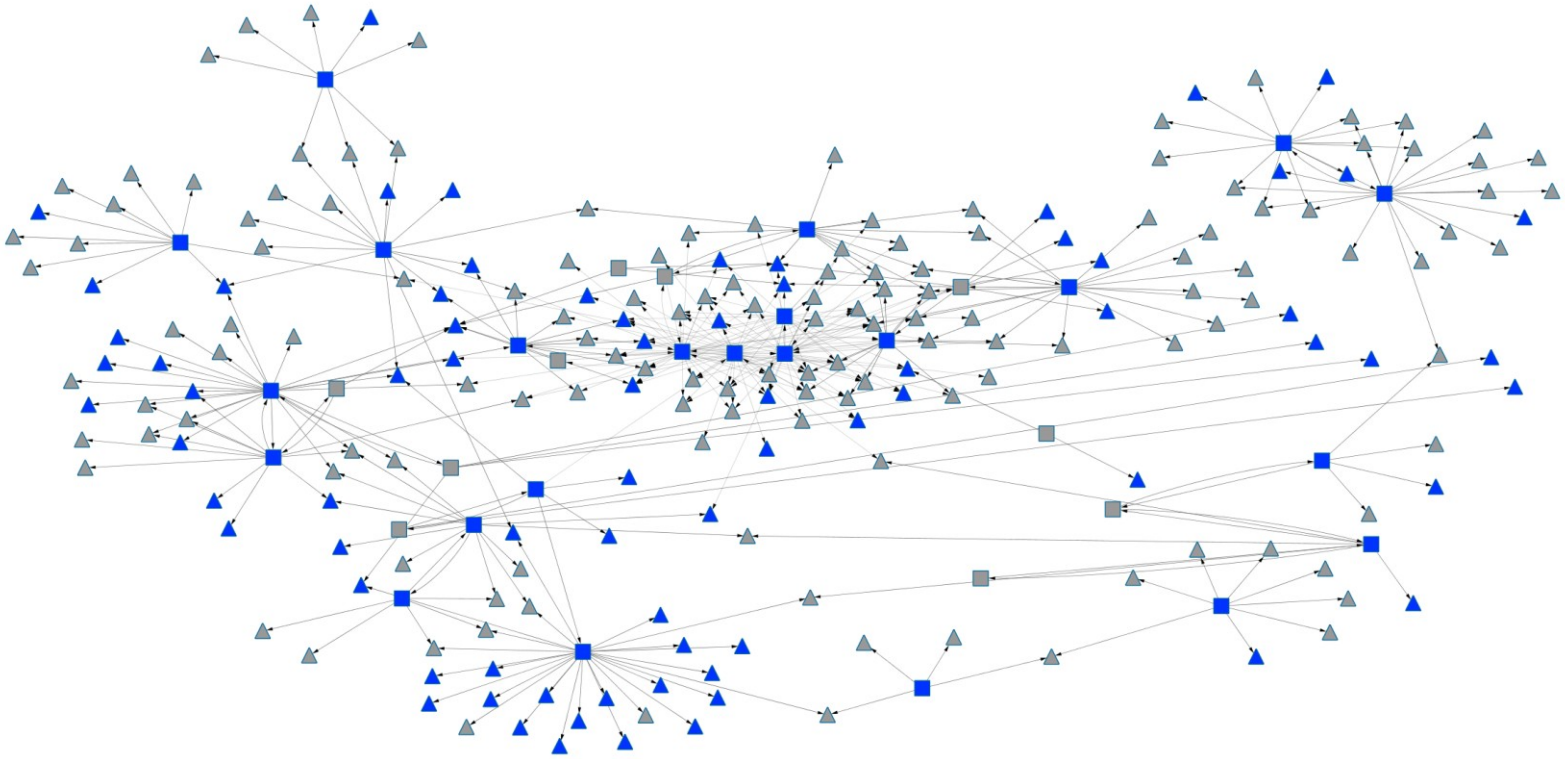
Node Shape	Classification
	Target Gene
	Transcription Factor

Pathogenic Motifs: Sparse Disconnect

- Exhibited in **age-related macular degeneration** (types 1a and 1b)
  - 4 diseases found with this motif
- Enrichment of high **indegree** (p = 0.0080)
- Enrichment of low **outdegree** (p = .
- Low density within pathogenic sub-network (p = .0161)
  - Pathogenic transcription factors and genes are disconnected
  - 25+ components



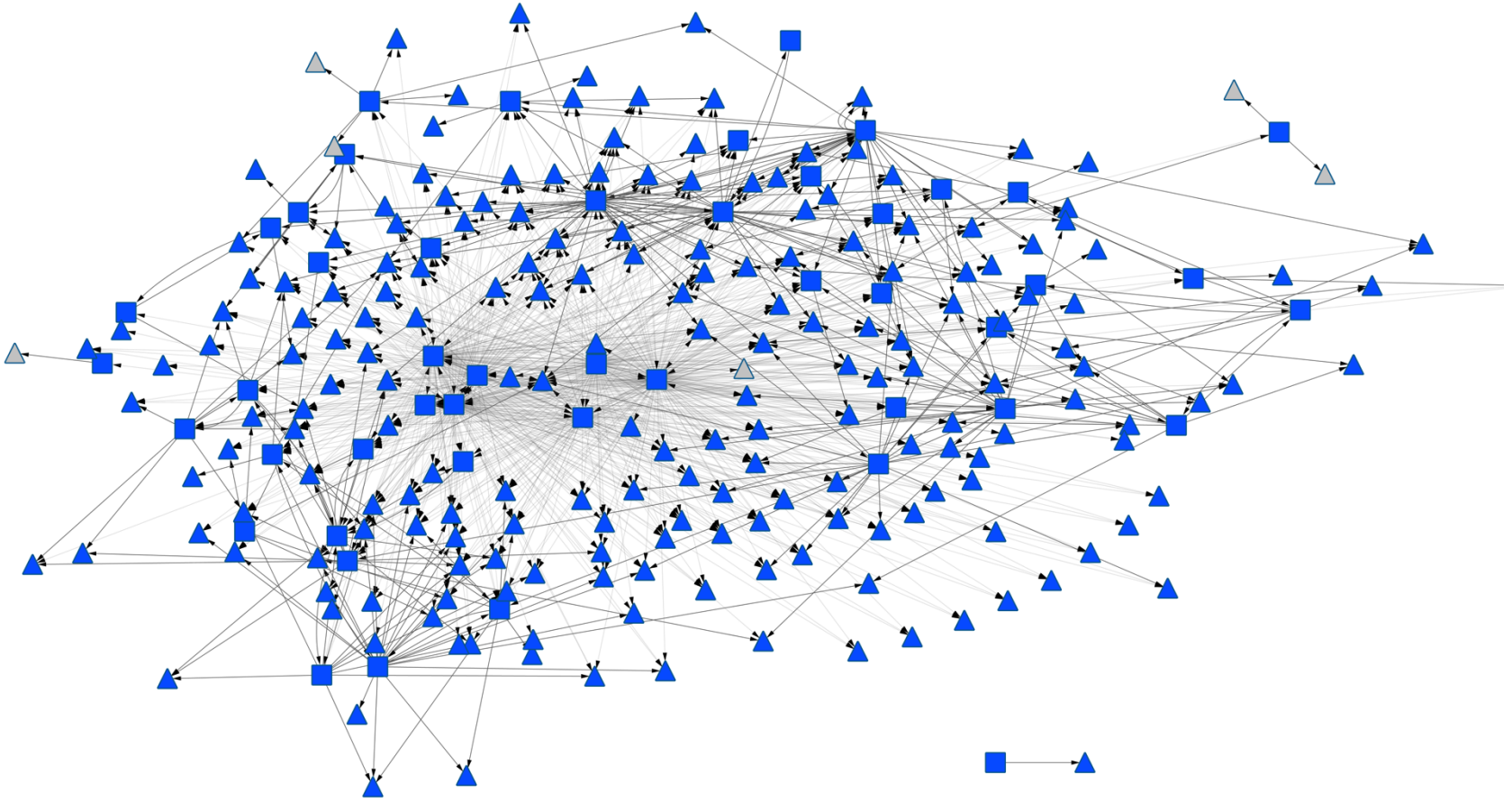
Sparse Disconnect Visualization



Node Fill Color	Pathogenicity
	Non-pathogenic (p > .01)
	Pathogenic (p < .01)

Node Shape	Classification
	Target Gene
	Transcription Factor

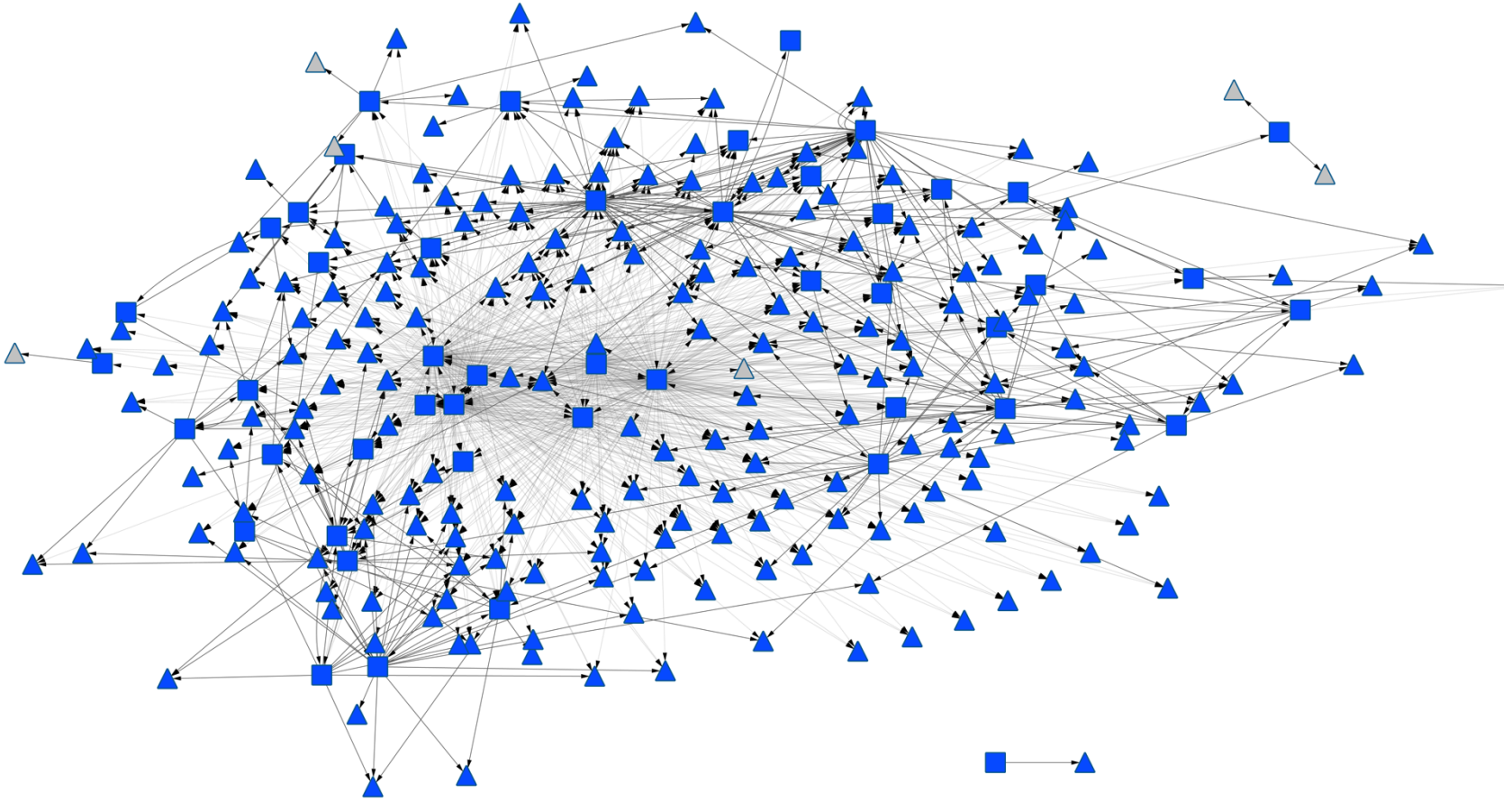
Low Distance Clustering Visualization



Pathogenic Motifs: Low Distance Clustering

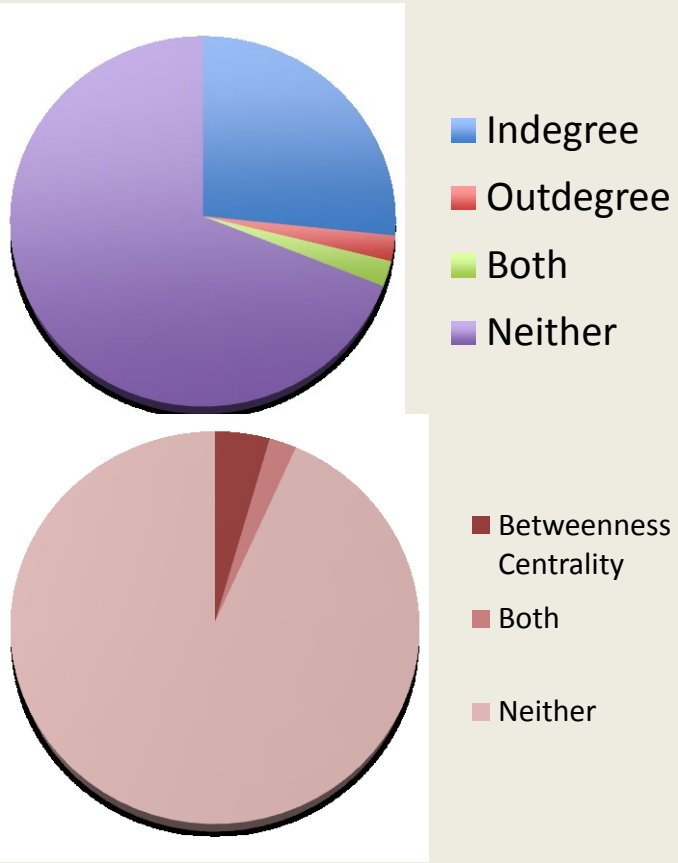
- Exhibited in **schizophrenia** (type 2)
- Enrichment for both high **indegree** ($p = .0084$) and high **outdegree** ($p = .0548$)
  - Positive feedback
- Enrichment for high **betweenness centrality** ($p = .0481$) and high **eigenvector centrality** ($p = .0605$)
- High density within pathogenic sub-network ($p = .0239$)
- 99% of genes are in a single connected component

Low Distance Clustering Visualization



Network Metrics

- Enrichment of indegree or outdegree was present in 36% of diseases
- Centrality measures were enriched in 9% of diseases
- No diseases were consistently enriched over the genes' clustering coefficient



Topics

- Background
  - Genetics
  - Regulatory Networks
  - The Human Regulatory Network
- Network Motifs
  - Questions and Methods
  - Sparse Disconnect
  - Low Distance Clustering
  - Network Metrics
- **Clustering Detection**
  - Method
  - Clusters Found

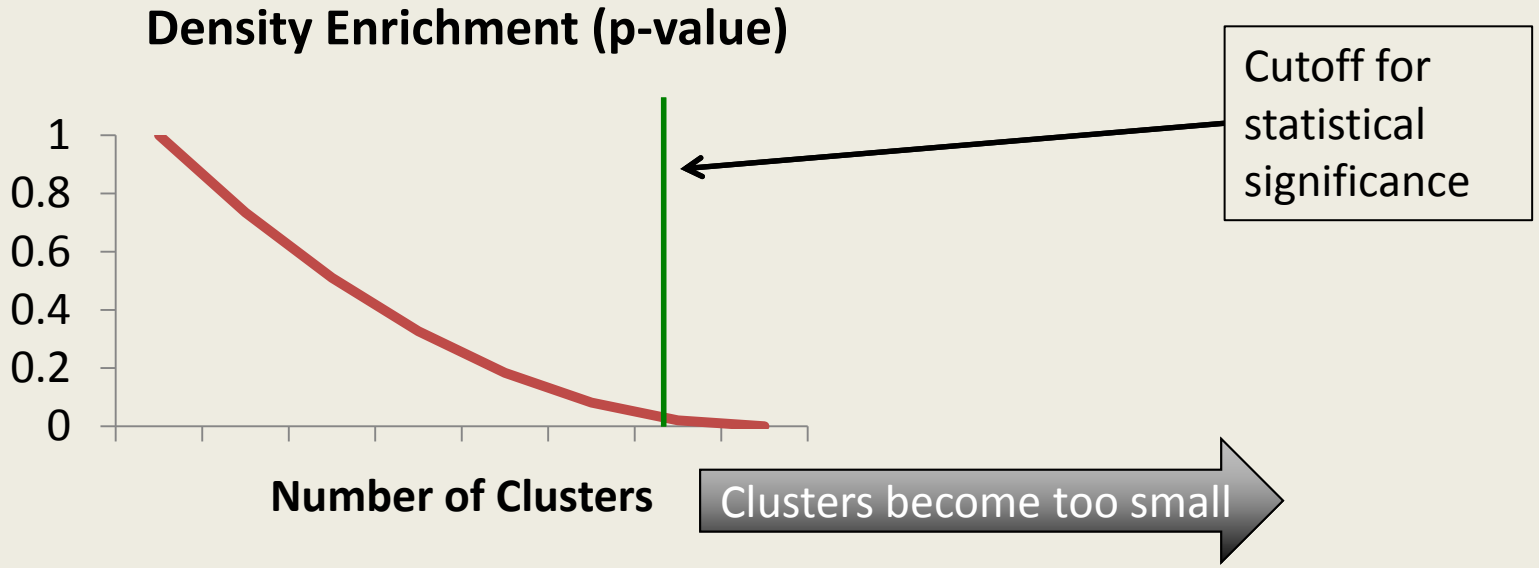
Clustering

- Another point of interest for genetic diseases
- Searching for cohesive regulatory units
- Provides more information about how the pathogenic genes interact

Cluster Detection

- Detects clusters through **spectral clustering**
  - Simplest form: uses network's algebraic connectivity to divide the nodes into two groups

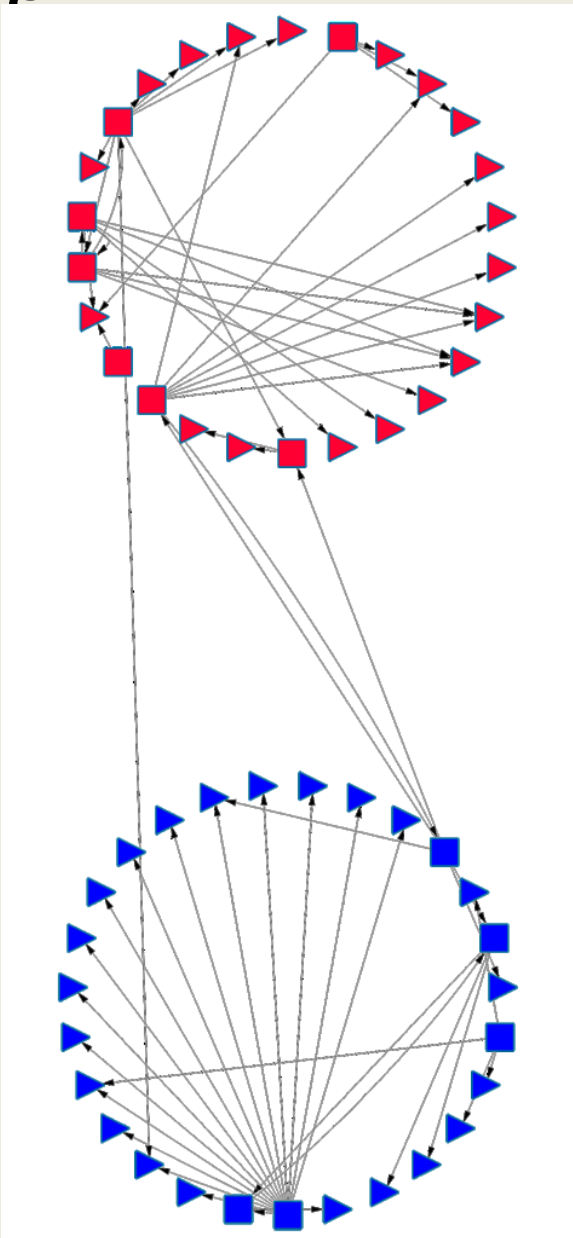
- Maximize cluster density and minimize cluster count



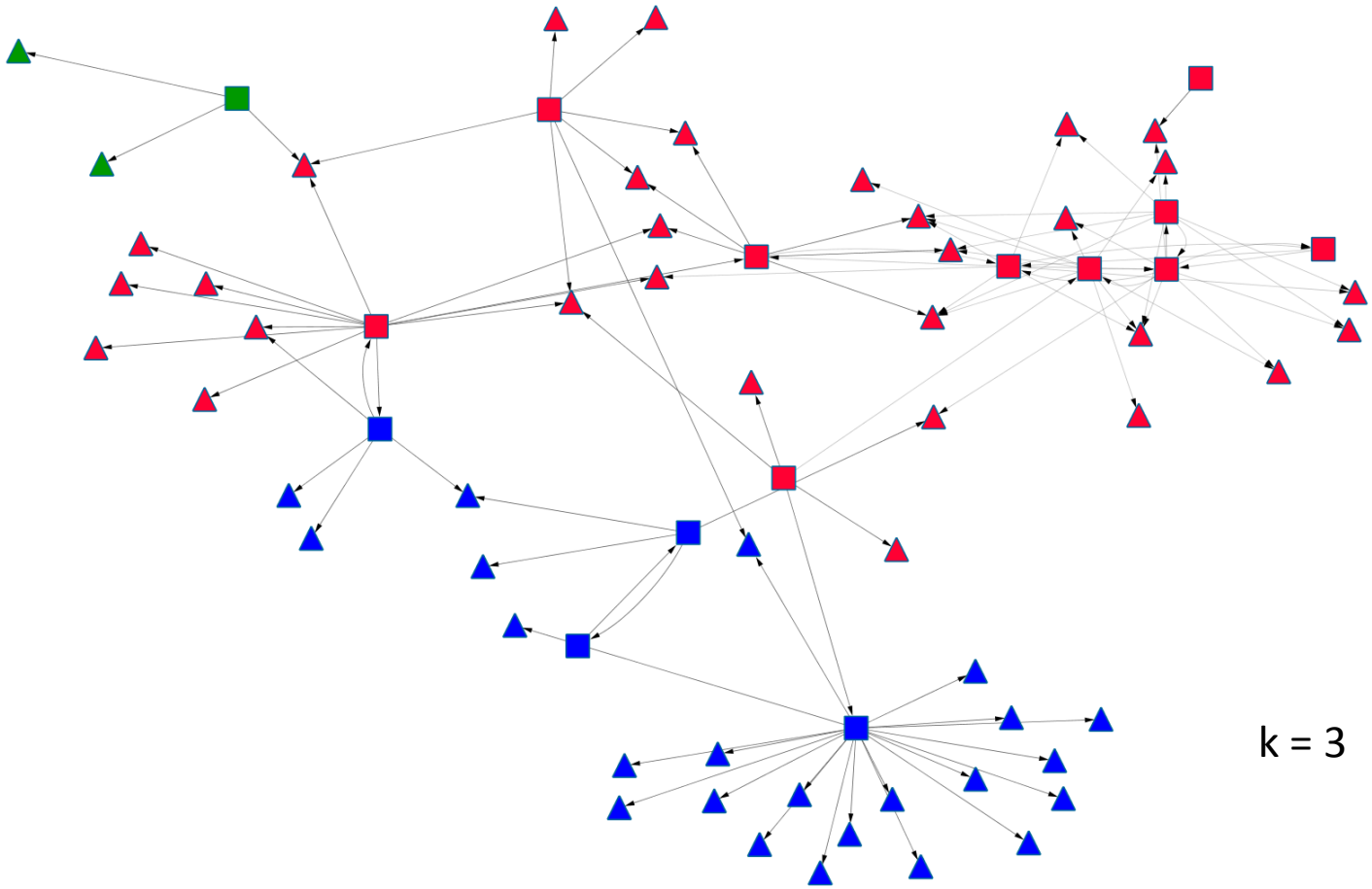
Spectral Clustering

- **Goal**: divide a network into two clusters such that the number of edges between *k* clusters is minimized

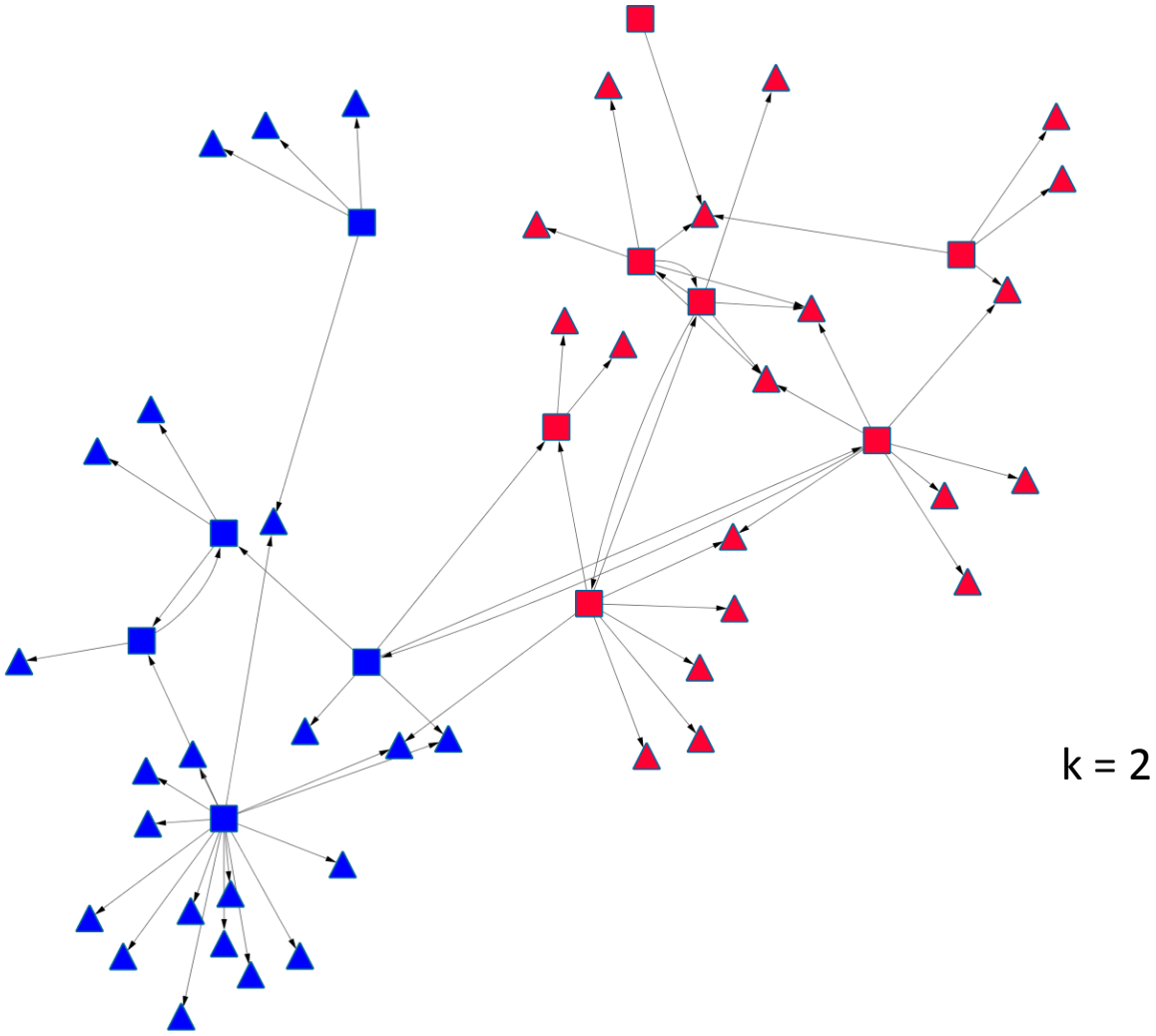
- **Method**: Combined spectral clustering with the k-means algorithm to optimize clusters



Age-related Macular Degeneration (type 1b) Clustering



Cardiovascular Disease Risk (type 1b) Clustering



Future Goals

- Continue search for pathogenic motifs
- Identify additional clusters
  - Different clustering algorithms
- Investigate GO terms within clusters

Thank You

- To **MIT PRIMES** for this engaging and challenging research opportunity

- To my mentor **Soheil Feizi** for his assistance and guidance throughout the project

- To **Professor Manolis Kellis** for suggesting the project