Graph Mining based clustering of PPI networks

Pallavi Gudipati Abhinav Garlapati

Computational Systems Biology

Outline

Objective

To cluster organisms based on the structure of their PPI networks.

Outline

Usage

- Can be used to analyze lesser known organisms.

Usage

- Can be used to analyze lesser known organisms.
- Basic information about the organism's PPI network can be derived from other members belonging to its cluster

Outline

PPI Networks

- STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a biological database and web resource of known and predicted protein-protein interactions.

PPI Networks

- STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a biological database and web resource of known and predicted protein-protein interactions.
- Networks considered:

Strain	Phylo group	Pathogenecity
E.coli O6:K15:H31 536	Group B2	UPEC
E.coli O103:H2 12009	Group B1	EHEC
E.coli C ATCC 8739	Group A	-
E.coli O127:H6 E2348/69	Group B2	EPEC

Outline

Background

Frequent Subgraph

Given a graph dataset $D = \{G_1, \cdots, G_n\}$ and a frequency threshold θ, a subgraph g with an observed support μ_0 is frequent if and only if $\mu_0 \geq \frac{\theta|D|}{100}$.

Background

Frequent Subgraph

Given a graph dataset $D = \{G_1, \cdots, G_n\}$ and a frequency threshold θ, a subgraph g with an observed support μ_0 is frequent if and only if $\mu_0 \geq \frac{\theta|D|}{100}$.

Maximal Subgraph

A subgaph g is a maximal subgraph in a graph dataset D if g is frequent, and there exists no supergraph $g^{'}$ such that $g \subset g^{'}$ and $g^{'}$ is frequent in D.

Background

Frequent Subgraph

Given a graph dataset $D = \{G_1, \cdots, G_n\}$ and a frequency threshold θ, a subgraph g with an observed support μ_0 is frequent if and only if $\mu_0 \geq \frac{\theta|D|}{100}$.

Maximal Subgraph

A subgaph g is a maximal subgraph in a graph dataset D if g is frequent, and there exists no supergraph $g^{'}$ such that $g \subset g^{'}$ and $g^{'}$ is frequent in D.

Statistical Significance

The statistical significance (or $p - value$) of a subgraph g with an observed support μ_0 is defined as the probability that it occurs in a random database with a support μ, where $\mu \geq \mu_0$.

Background: Random Walk with Restarts

- Random Walk simulates the trajectory taken by a random walker.

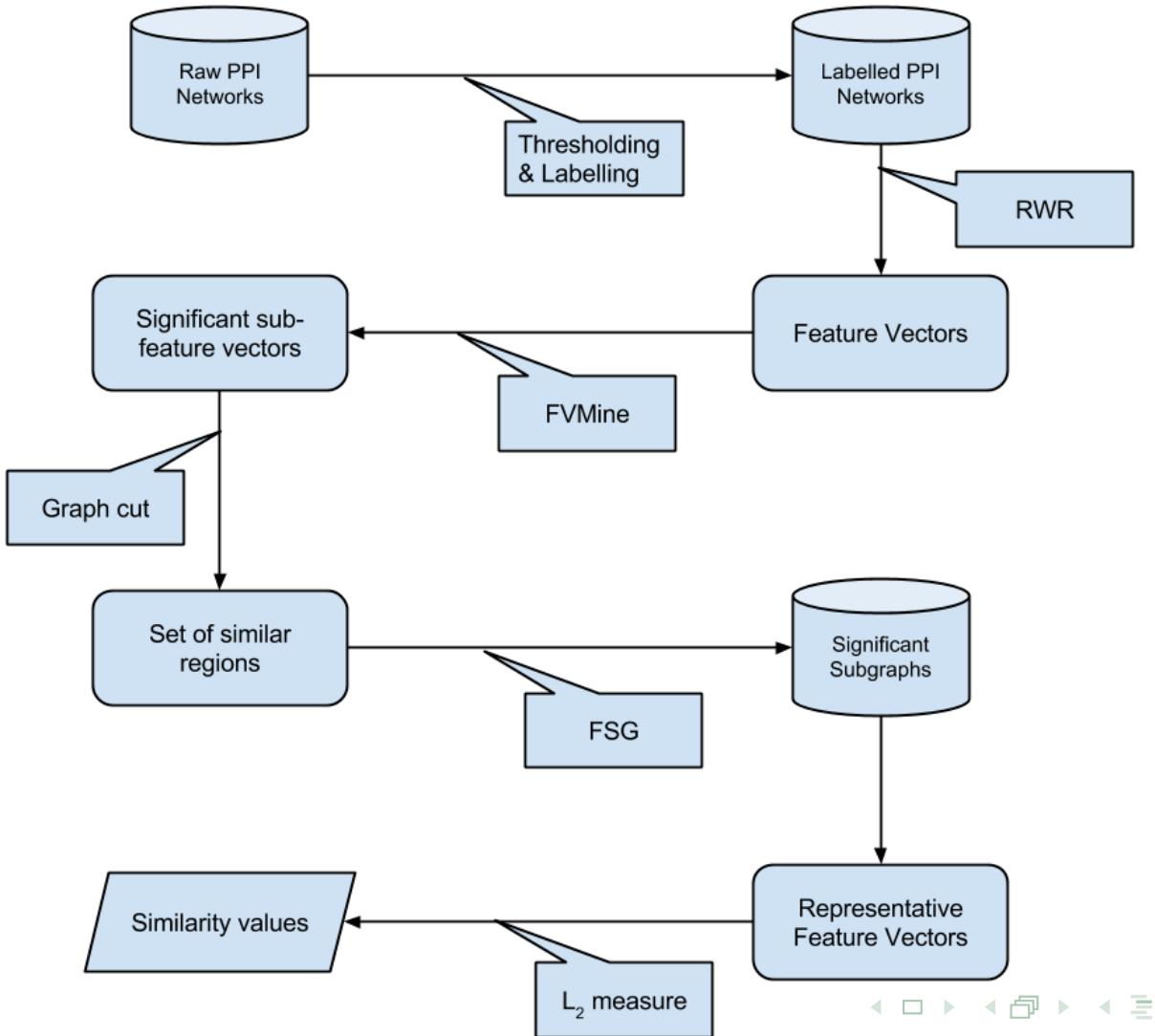
Background: Random Walk with Restarts

- Random Walk simulates the trajectory taken by a random walker.
- At each node, the walker can move along the edges originating from that node with equal probability.

Background: Random Walk with Restarts

- Random Walk simulates the trajectory taken by a random walker.
- At each node, the walker can move along the edges originating from that node with equal probability.
- To limit the neighbourhood in which the random walker can walk, we add a restart probability, β. Thus at every node, with certain probability the walker can be teleported back to the intial node.

Approach


Raw PPI Networks
Thresholding & Labelling
Labelled PPI Networks
RWR
Feature Vectors
FVMine
Significant sub-feature vectors
Graph cut
Set of similar regions
FSG
Significant Subgraphs
Representative Feature Vectors
L_2 measure
Similarity values

Outline

Naive Frequent Subgraph Mining Approach

- A frequent subgraph mining technique is used to mine all frequent subgraphs above a frequency threshold θ.

Naive Frequent Subgraph Mining Approach

- A frequent subgraph mining technique is used to mine all frequent subgraphs above a frequency threshold θ.
- Next, the $p - value$ (or some other significance measure) of each subgraph is computed and all subgraphs with a $p - value$ below a user-specified threshold are returned.

Naive Frequent Subgraph Mining Approach

- A frequent subgraph mining technique is used to mine all frequent subgraphs above a frequency threshold θ.
- Next, the $p - value$ (or some other significance measure) of each subgraph is computed and all subgraphs with a $p - value$ below a user-specified threshold are returned.
- But, we might miss significant subgraphs with low frequency.

Naive Frequent Subgraph Mining Approach

- A frequent subgraph mining technique is used to mine all frequent subgraphs above a frequency threshold θ.
- Next, the $p - value$ (or some other significance measure) of each subgraph is computed and all subgraphs with a $p - value$ below a user-specified threshold are returned.
- But, we might miss significant subgraphs with low frequency.
- **Solution**: We can lower θ but running time grows exponentially with decreasing frequency due to the inevitable explosion in graph search space.

Naive Frequent Subgraph Mining Approach

- A frequent subgraph mining technique is used to mine all frequent subgraphs above a frequency threshold θ.
- Next, the $p - value$ (or some other significance measure) of each subgraph is computed and all subgraphs with a $p - value$ below a user-specified threshold are returned.
- But, we might miss significant subgraphs with low frequency.
- **Solution**: We can lower θ but running time grows exponentially with decreasing frequency due to the inevitable explosion in graph search space.
- Need for a scalable technique to mine significant subgraphs from a graph database when the frequency of the subgraph is low.

Outline

Pre-processing

- Converted the weighted graph into an unweighted graph.

Pre-processing

- Converted the weighted graph into an unweighted graph.
- Labelling: Made node and edge types.

Outline

GraphSig

- Converted the neighborhood around each node to a feature vector using Random Walk with Restarts.

GraphSig

- Converted the neighborhood around each node to a feature vector using Random Walk with Restarts.
- Mined the generated feature vectors for significant and frequent sub-feature vectors. Vectors should meet the thresholds for both the $p - value$ and θ.

GraphSig

- Converted the neighborhood around each node to a feature vector using Random Walk with Restarts.
- Mined the generated feature vectors for significant and frequent sub-feature vectors. Vectors should meet the thresholds for both the $p - value$ and θ.
- Grouped all regions likely to contain a subgraph described by a significant sub-feature vector into sets.

GraphSig

- Converted the neighborhood around each node to a feature vector using Random Walk with Restarts.
- Mined the generated feature vectors for significant and frequent sub-feature vectors. Vectors should meet the thresholds for both the $p - value$ and θ.
- Grouped all regions likely to contain a subgraph described by a significant sub-feature vector into sets.
- Mined each set for maximal frequent subgraphs with a high frequency threshold.

Outline

Clustering

- Very few data points thus did not use any traditional clustering technique.

Clustering

- Very few data points thus did not use any traditional clustering technique.
- Converted each network to a representative feature vector.

Clustering

- Very few data points thus did not use any traditional clustering technique.
- Converted each network to a representative feature vector.
- Distance between any two networks is taken to be the Euclidean distance between their representative vectors.

Results

Table: L_2 Distances

Organism-1	Organism-2	Distance
E.coli O6:K15:H31 536	E.coli O103:H2 12009	66.7233092704
E.coli O6:K15:H31 536	E.coli C ATCC 8739	74.9466476902
E.coli O6:K15:H31 536	E.coli O127:H6 E2348/69	84.7761758987
E.coli O103:H2 12009	E.coli C ATCC 8739	22.2485954613
E.coli O103:H2 12009	E.coli O127:H6 E2348/69	45.3982378513
E.coli C ATCC 8739	E.coli O127:H6 E2348/69	48.3735464898

Results

- Overall 34 maximal frequent subgraphs were obtained.

Results

- Overall 34 maximal frequent subgraphs were obtained.
- The biggest subgraph obtained(node wise) had 7 nodes and 6 edges.

Results

- Overall 34 maximal frequent subgraphs were obtained.
- The biggest subgraph obtained(node wise) had 7 nodes and 6 edges.

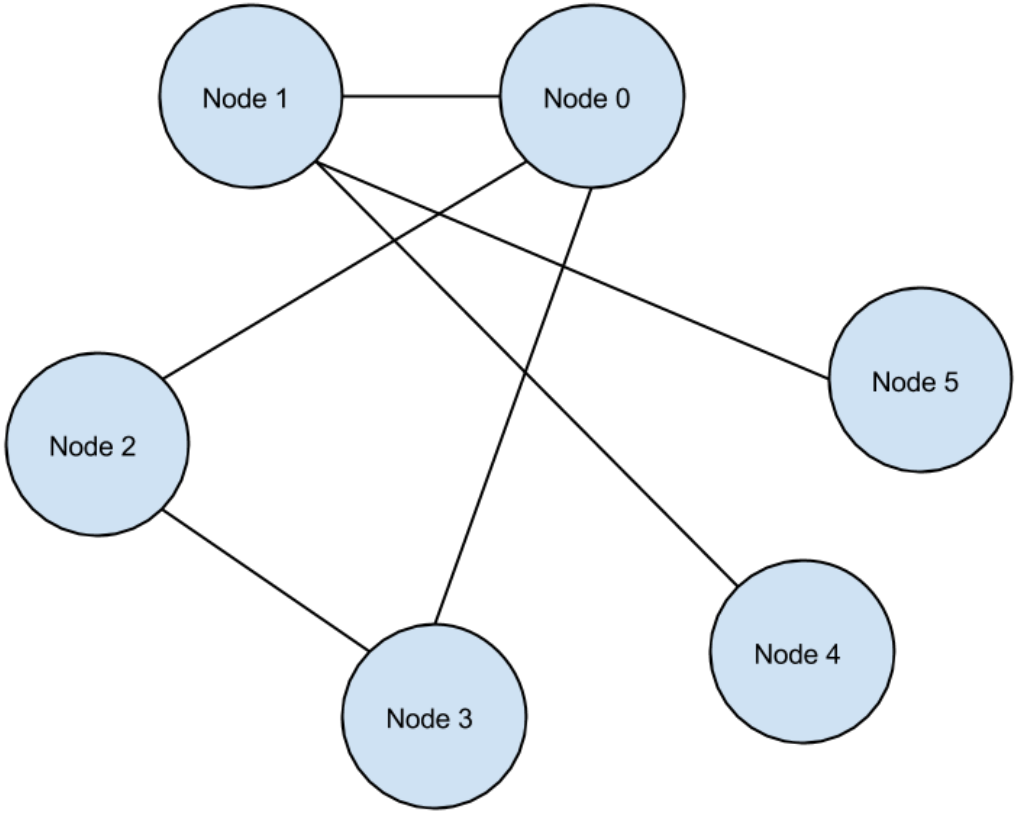


Figure: Maximal FS: 6 nodes, 6 edges

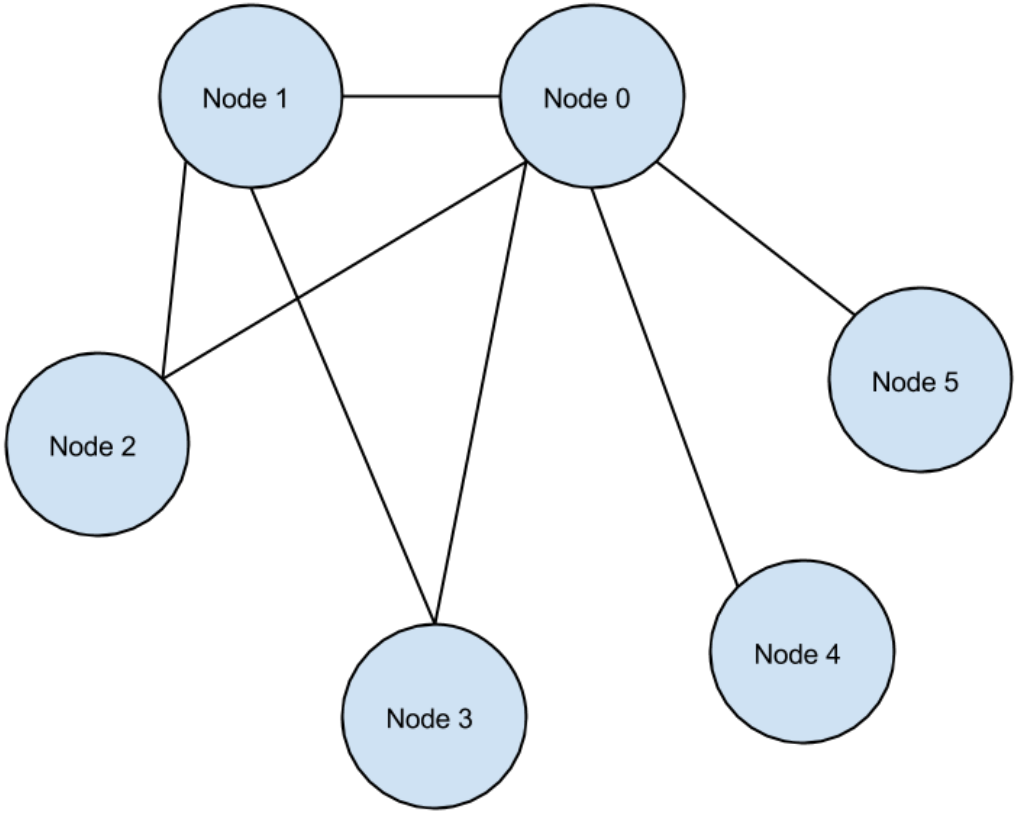


Figure: Maximal FS: 6 nodes, 7 edges

Observations

- Strains that are farthest apart also belong to the same phylogenetic group.

Observations

- Strains that are farthest apart also belong to the same phylogenetic group.
- The phylogeny of the organism might not affect the structure of the network, atleast in the way we are interpreting the networks. Maybe the functionality can have more correspondence.

Observations

- Strains that are farthest apart also belong to the same phylogenetic group.
- The phylogeny of the organism might not affect the structure of the network, atleast in the way we are interpreting the networks. Maybe the functionality can have more correspondence.
- If networks are very dissimilar, FVMine takes unreasonable amount of time to run. Also a skew is introduced in the significant subgraphs mined.

Observations

- Strains that are farthest apart also belong to the same phylogenetic group.
- The phylogeny of the organism might not affect the structure of the network, atleast in the way we are interpreting the networks. Maybe the functionality can have more correspondence.
- If networks are very dissimilar, FVMine takes unreasonable amount of time to run. Also a skew is introduced in the significant subgraphs mined.
- For highly significant nodes, almost the whole graph was reachable within 5 steps.

Future Work

- Attach more biological significance to the results.

Future Work

- Attach more biological significance to the results.
- Adapt this method to weighted graphs.

Future Work

- Attach more biological significance to the results.
- Adapt this method to weighted graphs.
- Introduce a regularizer to limit the number of significant graphs contributed by one organism to counter the skew introduced due to dissimilar networks.

Future Work

- Attach more biological significance to the results.
- Adapt this method to weighted graphs.
- Introduce a regularizer to limit the number of significant graphs contributed by one organism to counter the skew introduced due to dissimilar networks.
- Transition from Graph metrics to biological metrics:

Future Work

- Attach more biological significance to the results.
- Adapt this method to weighted graphs.
- Introduce a regularizer to limit the number of significant graphs contributed by one organism to counter the skew introduced due to dissimilar networks.
- Transition from Graph metrics to biological metrics:
  - Labelling of nodes

Future Work

- Attach more biological significance to the results.
- Adapt this method to weighted graphs.
- Introduce a regularizer to limit the number of significant graphs contributed by one organism to counter the skew introduced due to dissimilar networks.
- Transition from Graph metrics to biological metrics:
  - Labelling of nodes
  - Edge types