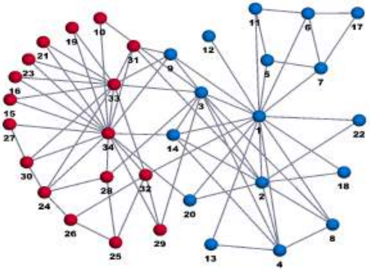
Functional Correlation Clustering with Applications in Cancer Genomics

Presenter: Gregory J. Puleo
Joint work with A. Emad, J. Hou, J. Ma and O. Milenkovic

Coordinated Science Laboratory
University of Illinois at Urbana-Champaign

2015 MBMC Workshop
University of Southern California
December 4, 2015

Community Detection and Recovery



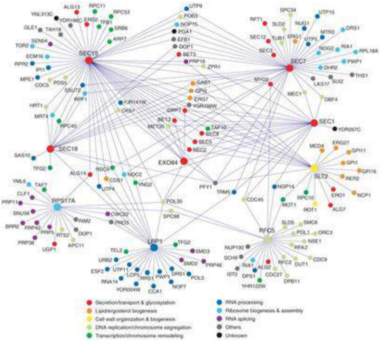
Zachary's karate club graph (Zachary 1977)

Image source:

`ifisc.uib-csic.es/jramasco/ComplexNets`

- Social networks: given pairwise connections (friendships) between individuals, identify "communities" using edge density information. (Undirected graphs)

Community Detection and Recovery


SEC15
SEC7
SEC1
SEC18
EXO84
SLT2
RPS17A
LRP1
RFC5
Secretion/transport & glycosylation
Lipid/ergosterol biogenesis
Cell wall organization & biogenesis
DNA replication/chromosome segregation
Transcription/chromosome remodeling
RNA processing
Ribosome biogenesis & assembly
RNA splicing
Others
Unknown

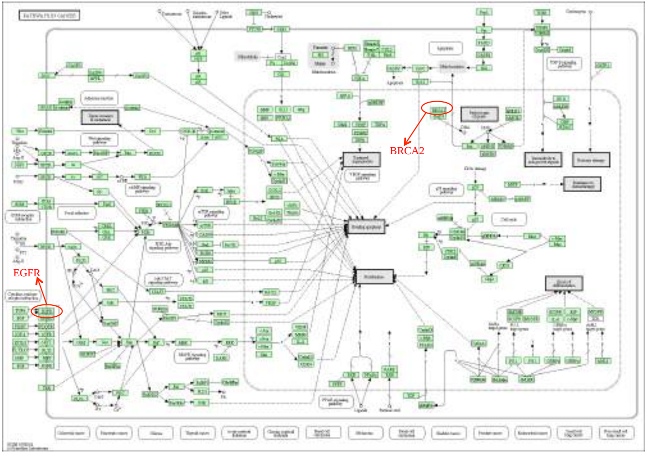
Gene regulatory networks (Davierwala et al. 2005)

- Social networks: given pairwise connections (friendships) between individuals, identify "communities" using edge density information. (Undirected graphs)

- Gene networks: given activating and suppressing interactions between genes, identify dense gene pathways. (Directed graphs)

Motivating Problem: Cancer Driver "Modules"

Significant recent interest in "cancer driver mutation" analysis. Driver mutations responsible for tumorigenesis, unlike neutral passenger mutations.

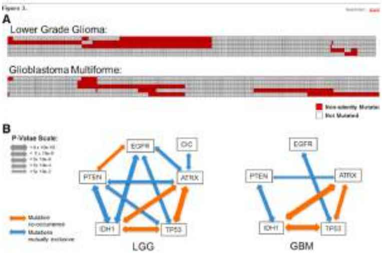


Motivating Problem: Cancer Driver “Modules”

- Significant recent interest in “cancer driver mutation” analysis. Driver mutations responsible for tumorigenesis, unlike neutral passenger mutations. Drivers obey:

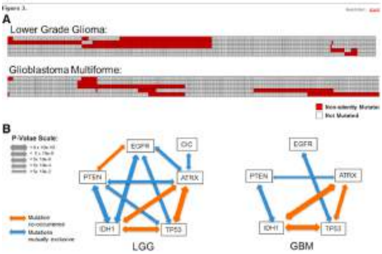
Motivating Problem: Cancer Driver “Modules”

- Significant recent interest in “cancer driver mutation” analysis. Driver mutations responsible for tumorigenesis, unlike neutral passenger mutations. Drivers obey:
- Mutual exclusivity principle: at most one mutation per pathway.



Motivating Problem: Cancer Driver “Modules”

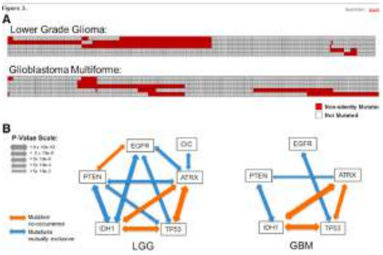
- Significant recent interest in “cancer driver mutation” analysis. Driver mutations responsible for tumorigenesis, unlike neutral passenger mutations. Drivers obey:
- Mutual exclusivity principle: at most one mutation per pathway.



- Pathway coverage property: Important pathways mutated in majority of patients.

Motivating Problem: Cancer Driver “Modules”

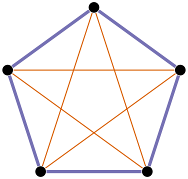
- Significant recent interest in “cancer driver mutation” analysis. Driver mutations responsible for tumorigenesis, unlike neutral passenger mutations. Drivers obey:
- Mutual exclusivity principle: at most one mutation per pathway.



- Pathway coverage property: Important pathways mutated in majority of patients.
- Pathway constraints: Number of driver pathways unknown, sizes “mostly” limited to 6-15 genes.

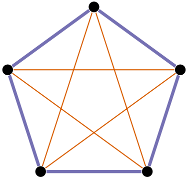
Functional Community Detection: Correlation Clustering

- Correlation clustering (clustering with qualitative information) was introduced by Bansal, Blum, and Chawla (FOCS '02).



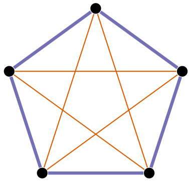
Functional Community Detection: Correlation Clustering

- Correlation clustering (clustering with qualitative information) was introduced by Bansal, Blum, and Chawla (FOCS '02).
- Basic formulation: one is given n objects and, for some pairs of objects, a judgment whether they are **similar** or dissimilar. Similarity may be assessed based on given data.



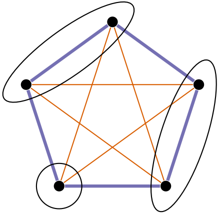
Functional Community Detection: Correlation Clustering

- Correlation clustering (clustering with qualitative information) was introduced by Bansal, Blum, and Chawla (FOCS '02).
- Basic formulation: one is given n objects and, for some pairs of objects, a judgment whether they are **similar** or dissimilar. Similarity may be assessed based on given data.
- Represent constraints via graph G with edges labeled $+$ or $-$.



Functional Community Detection: Correlation Clustering

- Correlation clustering (clustering with qualitative information) was introduced by Bansal, Blum, and Chawla (FOCS '02).
- Basic formulation: one is given n objects and, for some pairs of objects, a judgment whether they are **similar** or dissimilar. Similarity may be assessed based on given data.
- Represent constraints via graph G with edges labeled $+$ or $-$.
- Goal: Partition vertices into clusters by minimizing total number of $+$ edges between clusters and the number of $-$ edges within clusters.



Correlation Clustering: Basics

- Correlation Clustering does not require advance knowledge of number of clusters, and is a form of "agnostic learning."

Correlation Clustering: Basics

- Correlation Clustering does not require advance knowledge of number of clusters, and is a form of "agnostic learning."
- Related to cluster editing, graph partitioning, sparsest cut, balanced edge partitioning, rank aggregation...

Correlation Clustering: Basics

- Correlation Clustering does not require advance knowledge of number of clusters, and is a form of "agnostic learning."
- Related to cluster editing, graph partitioning, sparsest cut, balanced edge partitioning, rank aggregation...
- Finding best clustering is NP-Hard (Bansal–Blum–Chawla), so we seek approximation algorithms.

Correlation Clustering: Basics

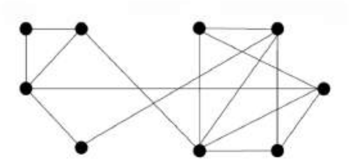
- Correlation Clustering does not require advance knowledge of number of clusters, and is a form of "agnostic learning."
- Related to cluster editing, graph partitioning, sparsest cut, balanced edge partitioning, rank aggregation...
- Finding best clustering is NP-Hard (Bansal–Blum–Chawla), so we seek approximation algorithms.
- If underlying graph can have missing edges, probably no constant-factor approximation algorithm is possible. Thus, we assume the underlying graph is complete.

Correlation Clustering: Basics

- Correlation Clustering does not require advance knowledge of number of clusters, and is a form of "agnostic learning."
- Related to cluster editing, graph partitioning, sparsest cut, balanced edge partitioning, rank aggregation...
- Finding best clustering is NP-Hard (Bansal–Blum–Chawla), so we seek approximation algorithms.
- If underlying graph can have missing edges, probably no constant-factor approximation algorithm is possible. Thus, we assume the underlying graph is complete.
- We discuss two algorithms: Ailon–Charikar–Newman (2008) and Charikar–Guruswami–Wirth (2005).

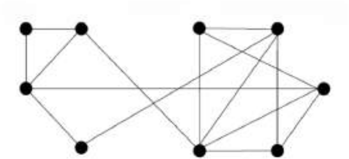
Algorithms: ACN Pivoting

- Start with complete graph with negative edges removed:

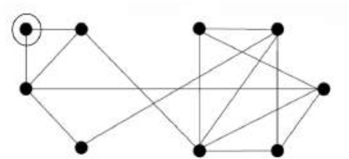


Algorithms: ACN Pivoting

- Start with complete graph with negative edges removed:

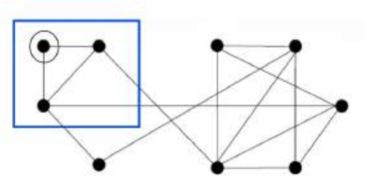


- Choose vertex uniformly at random:



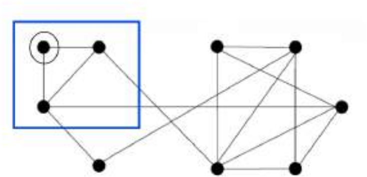
Algorithms: ACN Pivoting

- Form cluster from its neighborhood:

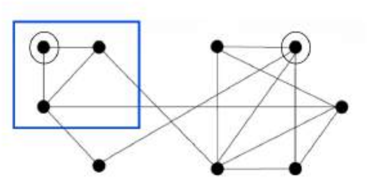


Algorithms: ACN Pivoting

- Form cluster from its neighborhood:

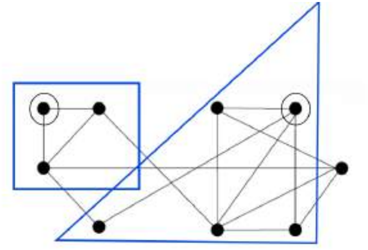


- Choose another vertex uniformly at random:



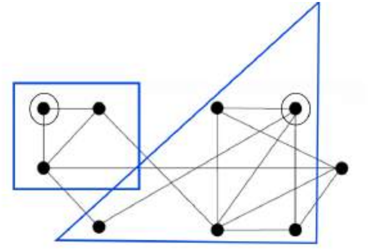
Algorithms: ACN Pivoting

- Form cluster from its neighborhood:



Algorithms: ACN Pivoting

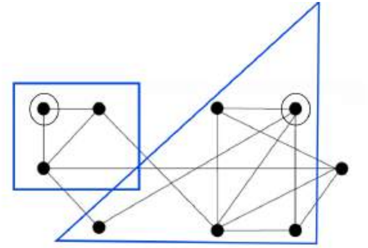
- Form cluster from its neighborhood:



- Continue.

Algorithms: ACN Pivoting

- Form cluster from its neighborhood:



- Continue.
- This yields a 3-approximation algorithm; can improve to 2.5 by solving an LP first.

Algorithms: CGW, Step 1 – LP Relaxation

Interpret $x_{uv} = 1$ as saying "u and v are in different clusters", interpret $x_{uv} = 0$ as saying "u and v are in the same cluster".

Solve LP relaxation of integer formulation of correlation clustering:

$$\begin{array}{ll} \underset{0 \leq x_e \leq 1}{\text{minimize}} & \displaystyle\sum_{e \in \boldsymbol{E}^{+}(G)} x_e + \sum_{e \in E^{-}(G)} (1 - x_e) \\ \text{subject to} & x_{uv} \leq x_{uz} + x_{zv} \quad (\forall \text{ distinct } u, v, z \in V(G)) \end{array}$$

Cost of optimal LP solution $\leq$ cost of optimal clustering.

Algorithms: CGW, Step 2 – Rounding Up

Round up LP solution x to obtain clustering (region growing):

```
Let $S = V(G)$.
while $S \neq \emptyset$ do
    Let the “pivot vertex” $u$ be an arbitrary element of $S$.
    Let $T = \{w \in S - \{u\} \colon x_{uw} \leq 1/2\}$.
    if $\sum_{w \in T} x_{uw} \geq |T|/4$ then
        Output the singleton cluster $\{u\}$.
        Let $S = S - \{u\}$.
    else
        Output the cluster $\{u\} \cup T$.
        Let $S = S - (\{u\} \cup T)$.
    end if
end while
```

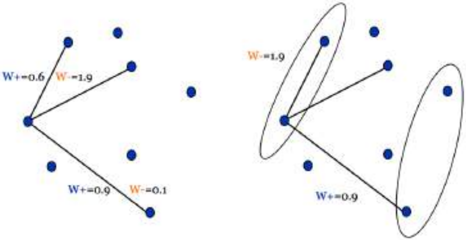
Cost of resulting clustering is $\leq$ 4 times cost of optimal clustering.

Extensions of the Basic Model: Edge Weights

- Ailon–Charikar–Newman also introduced *weighted* formulation: instead of $+/-$ labels, each edge e has weights $\boldsymbol{w_e^+}, w_e^- \geq 0$.

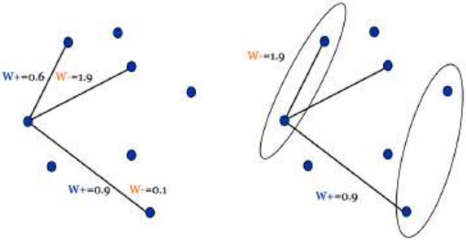
Extensions of the Basic Model: Edge Weights

- Ailon–Charikar–Newman also introduced *weighted* formulation: instead of $+/-$ labels, each edge e has weights $w_e^+, w_e^- \geq 0$.
- Edge e costs w_e^+ when placed between clusters and costs w_e^- when placed within a cluster.



Extensions of the Basic Model: Edge Weights

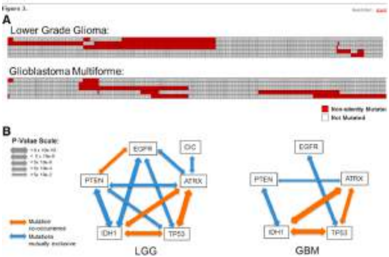
- Ailon–Charikar–Newman also introduced *weighted* formulation: instead of $+/-$ labels, each edge e has weights $\boldsymbol{w_e^+}, w_e^- \geq 0$.
- Edge e costs $\boldsymbol{w_e^+}$ when placed between clusters and costs w_e^- when placed within a cluster.



- Known: if all $\boldsymbol{w_e^+} + w_e^- = 1$ ("probability constraints"), there is a randomized expected 2.5-approximation algorithm (Ailon-Charikar-Newman).

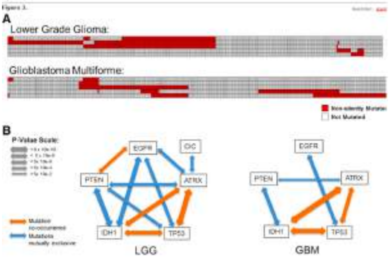
Recall Motivating Problem: Cancer Driver Modules

- Mutual exclusivity principle: at most one mutation per pathway.



Recall Motivating Problem: Cancer Driver Modules

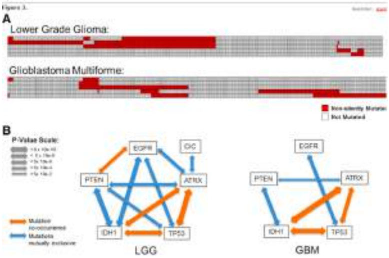
- Mutual exclusivity principle: at most one mutation per pathway.



- Pathway coverage property: Important pathways mutated in majority of patients.

Recall Motivating Problem: Cancer Driver Modules

- Mutual exclusivity principle: at most one mutation per pathway.



- Pathway coverage property: Important pathways mutated in majority of patients.
- Pathways constraints: Number unknown, sizes “mostly” limited to 6-15 genes.

Extensions: Size Constraints and New Weights

- Need to cover larger set of allowed weights and discourage “large” clusters in solution. In latter case, introduce a size bound parameter K, and give each vertex v a parameter $\mu_v \geq 0$.

Extensions: Size Constraints and New Weights

- Need to cover larger set of allowed weights and discourage "large" clusters in solution. In latter case, introduce a size bound parameter K, and give each vertex v a parameter $\mu_v \geq 0$.
- If v is in a cluster of size $K + 1 + y_v$, where $y_v \geq 0$, then vertex v is assigned a "penalty" of $\mu_v y_v$.

Extensions: Size Constraints and New Weights

- Need to cover larger set of allowed weights and discourage “large” clusters in solution. In latter case, introduce a size bound parameter K, and give each vertex v a parameter $\mu_v \geq 0$.
- If v is in a cluster of size $K + 1 + y_v$, where $y_v \geq 0$, then vertex v is assigned a “penalty” of $\mu_v y_v$.
- Setting all $\mu_v = 0$ gives no size bounds.

Extensions: Size Constraints and New Weights

- Need to cover larger set of allowed weights and discourage "large" clusters in solution. In latter case, introduce a size bound parameter K, and give each vertex v a parameter $\mu_v \geq 0$.
- If v is in a cluster of size $K + 1 + y_v$, where $y_v \geq 0$, then vertex v is assigned a "penalty" of $\mu_v y_v$.
- Setting all $\mu_v = 0$ gives no size bounds.
- If $\mu_v \geq \max_e \boldsymbol{w_e^+}$, size constraints are "hard": always profitable to split an oversized cluster.

Extensions: Size Constraints and New Weights

- Need to cover larger set of allowed weights and discourage "large" clusters in solution. In latter case, introduce a size bound parameter K, and give each vertex v a parameter $\mu_v \geq 0$.
- If v is in a cluster of size $K + 1 + y_v$, where $y_v \geq 0$, then vertex v is assigned a "penalty" of $\mu_v y_v$.
- Setting all $\mu_v = 0$ gives no size bounds.
- If $\mu_v \geq \max_e \boldsymbol{w_e^+}$, size constraints are "hard": always profitable to split an oversized cluster.
- Intermediate values yield "soft constraints", where we might allow a large cluster if it has very few errors otherwise.

Extensions: Size Constraints and New Weights

- Need to cover larger set of allowed weights and discourage "large" clusters in solution. In latter case, introduce a size bound parameter K, and give each vertex v a parameter $\mu_v \geq 0$.
- If v is in a cluster of size $K + 1 + y_v$, where $y_v \geq 0$, then vertex v is assigned a "penalty" of $\mu_v y_v$.
- Setting all $\mu_v = 0$ gives no size bounds.
- If $\mu_v \geq \max_e \boldsymbol{w_e^+}$, size constraints are "hard": always profitable to split an oversized cluster.
- Intermediate values yield "soft constraints", where we might allow a large cluster if it has very few errors otherwise.
- Deal with all these difficulties by extending CGW algorithm.

The Bounded Cluster Size Algorithm: LP Part

To accommodate weights and size constraints, modify the CGW LP as:

$$
\begin{array}{lll}
\underset{\substack{0 \le x_e \le 1 \\ y_v \ge 0}}{\text{minimize}} & \left[\sum_{e \in E(G)} \left(\boldsymbol{w_e^+} x_e + w_e^- (1 - x_e) \right) \right] + \sum_{v \in V(G)} \mu_v y_v & \\
\text{subject to} & x_{uv} \le x_{uz} + x_{zv} & (\forall \text{ distinct } u, v, z \in V(G)) \\
 & \sum_{u \ne v} (1 - x_{uv}) \le K + y_v & (\text{for all } v \in V(G))
\end{array}
$$

The Bounded Cluster Size Algorithm: Rounding Up Part

Modify rounding up algorithm by introducing new parameter α:

```
Let $S = V(G)$.
while $S \neq \emptyset$ do
    Let the "pivot vertex" $u$ be an arbitrary element of $S$.
    Let $T = \{w \in S - \{u\} \colon x_{uw} \leq \alpha\}$.
    if $\sum_{w \in T} x_{uw} \geq \alpha \left| T \right| / 2$ then
        Output the singleton cluster $\{u\}$.
        Let $S = S - \{u\}$.
    else
        Output the cluster $\{u\} \cup T$.
        Let $S = S - (\{u\} \cup T)$.
    end if
end while
```

Original CGW algorithm uses $\alpha = 1/2$. Need to choose α to optimize approximation ratio, based on $\boldsymbol{w_e^+}$, w_e^-, y_v.

Correlation Clustering: Results

- First correlation clustering model with bounded cluster sizes, minimax and thresholding properties. Accompanying approximation algorithms based on LP relaxations and pivoting.

Correlation Clustering: Results

- First correlation clustering model with bounded cluster sizes, minimax and thresholding properties. Accompanying approximation algorithms based on LP relaxations and pivoting.

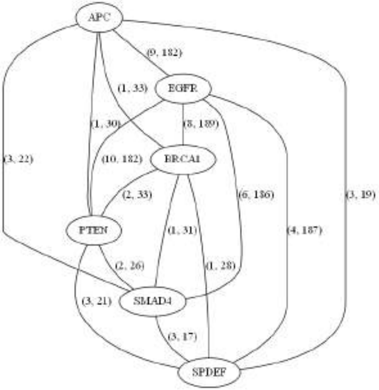
- Broadest range of weight values for which constant approximation algorithms provably exist for unconstrained, soft and hard constrained cluster sizes.

Correlation Clustering: Results

- First correlation clustering model with bounded cluster sizes, minimax and thresholding properties. Accompanying approximation algorithms based on LP relaxations and pivoting.
- Broadest range of weight values for which constant approximation algorithms provably exist for unconstrained, soft and hard constrained cluster sizes.
- State-of-the-art pathway discovery method based on mutual exclusivity analysis.

Correlation Clustering for Driver Module Analysis

Will construct complete, weighted graph G with vertices V indexed by genes, and edges labeled by weight vectors capturing:

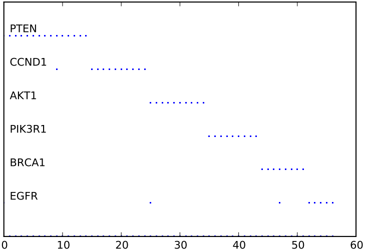


EGFR/PI3K/PTEN/Akt/mTORC1 driver group.

- Mutual exclusivity: handled via large negative weights for edges between genes that are rarely co-mutated.

Correlation Clustering for Driver Module Analysis

Will construct complete, weighted graph G with vertices V indexed by genes, and edges labeled by weight vectors capturing:



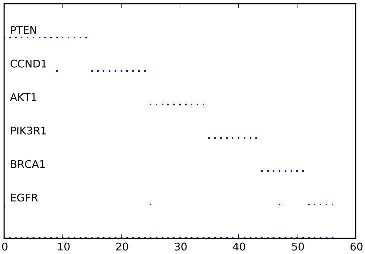
EGFR/PI3K/PTEN/Akt/mTORC1 driver group:

x-axis provides number of patients, y-axis lists genes.

- Mutual exclusivity: handled via large negative weights for edges between genes that are rarely co-mutated.
- Coverage: handled via weights of positive edges. May also include partial knowledge about network through Kegg database.

Correlation Clustering for Driver Module Analysis

Will construct complete, weighted graph G with vertices V indexed by genes, and edges labeled by weight vectors capturing:



EGFR/PI3K/PTEN/Akt/mTORC1 driver group:

x-axis provides number of patients, y-axis lists genes.

- Mutual exclusivity: handled via large negative weights for edges between genes that are rarely co-mutated.
- Coverage: handled via weights of positive edges. May also include partial knowledge about network through Kegg database.
- Few pathways: enforced via soft cluster size bounds.

Cancer Modules: Choosing the Weights

- Parameters: Number of patients n_p and number of genes $n_g = |V(G)|$. For each gene (i.e. vertex) u in G, $\mathcal{S}(u)$ denotes the set of patients in which u is mutated.

Cancer Modules: Choosing the Weights

- Parameters: Number of patients n_p and number of genes $n_g = |V(G)|$. For each gene (i.e. vertex) u in G, $\mathcal{S}(u)$ denotes the set of patients in which u is mutated.
- TCGA Breast Cancer Dataset: Total of $n_p = 504$ patients; $n_g = 8726$ genes.

Cancer Modules: Choosing the Weights

- Parameters: Number of patients n_p and number of genes $n_g = |V(G)|$. For each gene (i.e. vertex) u in G, $\mathcal{S}(u)$ denotes the set of patients in which u is mutated.
- TCGA Breast Cancer Dataset: Total of $n_p = 504$ patients; $n_g = 8726$ genes.
- Negative weights: For any $u, v \in V(G)$,

$$w^{-}_{u,v} = a \times \frac{|\mathcal{S}(u) \cap \mathcal{S}(v)|}{\min(|\mathcal{S}(u)|, |\mathcal{S}(v)|)},$$

where a is a user-specified relevance parameter.

Results on Cancer Modules: Choosing the Weights

- Positive weights: If two genes increase coverage significantly, their positive weight should be large and encourage placement in the same cluster.
  Let $\mathcal{D} = \{D(u,v)\}, \forall\ u, v \in V(G)$, where $D(u,v) = |\mathcal{S}(u)\ \Delta\ \mathcal{S}(v)|$. Let $T(J)$ be the Jth percentile of the values in $\mathcal{D}$. Then

$$\boldsymbol{w}_{uv}^{+} = \begin{cases} 1 & \text{if} \quad D(u,v) > T(J) \\ \frac{1}{T(J)} \times D(u,v) & \text{otherwise.} \end{cases}$$

Results on Cancer Modules: Choosing the Weights

- Positive weights: If two genes increase coverage significantly, their positive weight should be large and encourage placement in the same cluster.
  Let $\mathcal{D} = \{D(u,v)\}$, $\forall\ u, v \in V(G)$, where $D(u,v) = |\mathcal{S}(u)\ \Delta\ \mathcal{S}(v)|$. Let $T(J)$ be the Jth percentile of the values in $\mathcal{D}$. Then

$$\boldsymbol{w}^{+}_{uv} = \begin{cases} 1 & \text{if} \quad D(u,v) > T(J) \\ \frac{1}{T(J)} \times D(u,v) & \text{otherwise.} \end{cases}$$

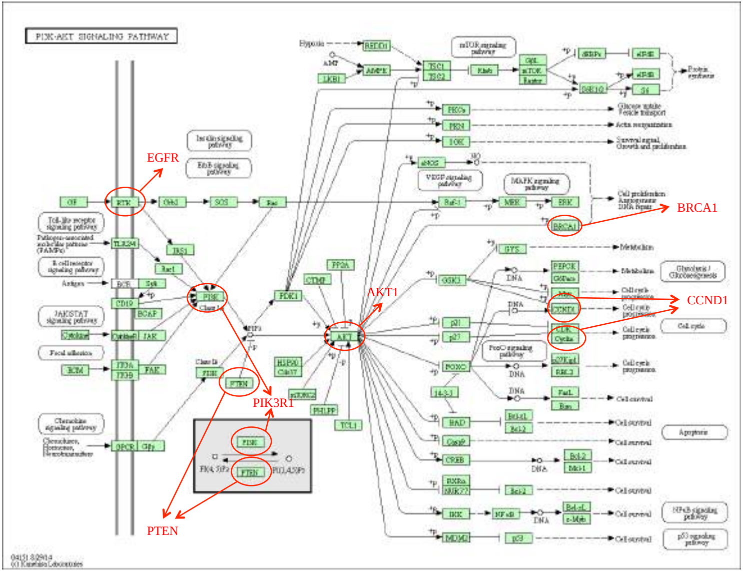
- To ensure $w^{-}_{uv} + \boldsymbol{w}^{+}_{uv} \geq 1$, for all $u, v \in V(G)$, also need

$$\text{If} \quad \boldsymbol{w}^{+}_{uv} + w^{-}_{uv} < 1,$$

$$\text{Set} \quad w^{-}_{uv} = \frac{w^{-}_{uv}}{\boldsymbol{w}^{+}_{uv} + w^{-}_{uv}},$$

$$\text{Set} \quad \boldsymbol{w}^{+}_{uv} = 1 - w^{-}_{uv}.$$

Results on Modules: Uniform Modules Within a Pathway



Thank You!

Questions?