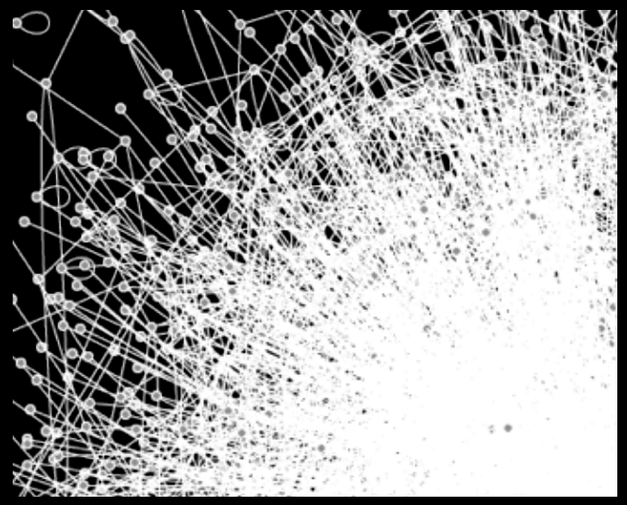


Clustering Uncertain Biological Networks

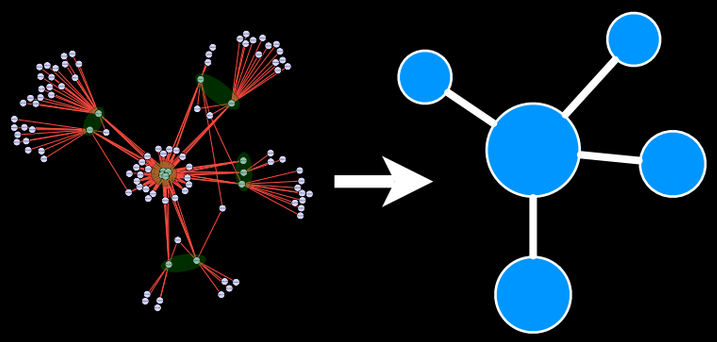
Carl Kingsford
Assistant Professor
Department of Computer Science
Center for Bioinformatics & Computational Biology
University of Maryland, College Park

Joint work with Saket Navlakha & David E. Kelley

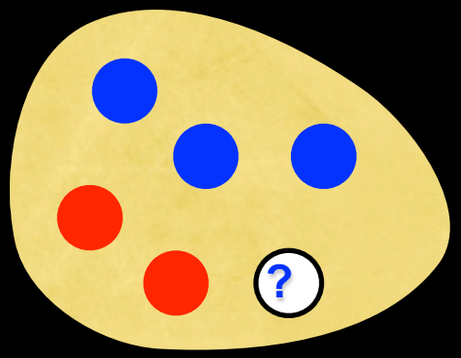
Why Cluster Biological Networks?

Networks based on protein binding, gene regulation, gene redundancy, & metabolic reactions are naturally “modular.”

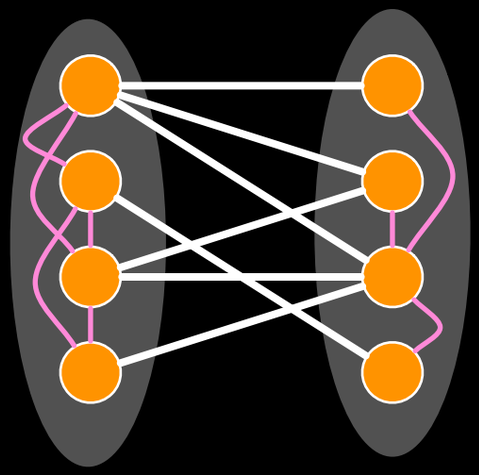
Summarize complexity, pattern mining, visualization



Associate proteins with diseases, functions, or complexes

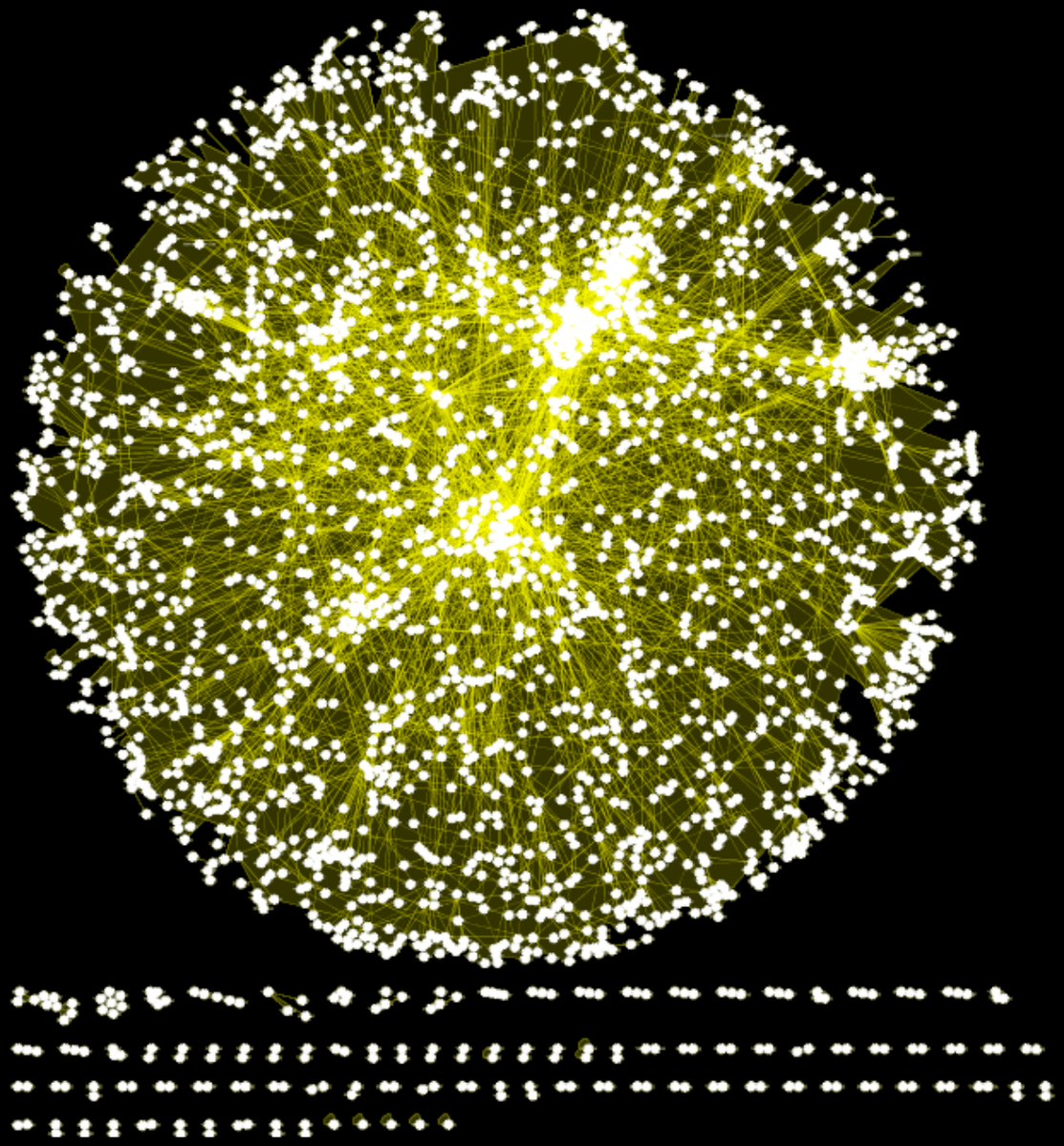


Uncover redundant cellular pathways

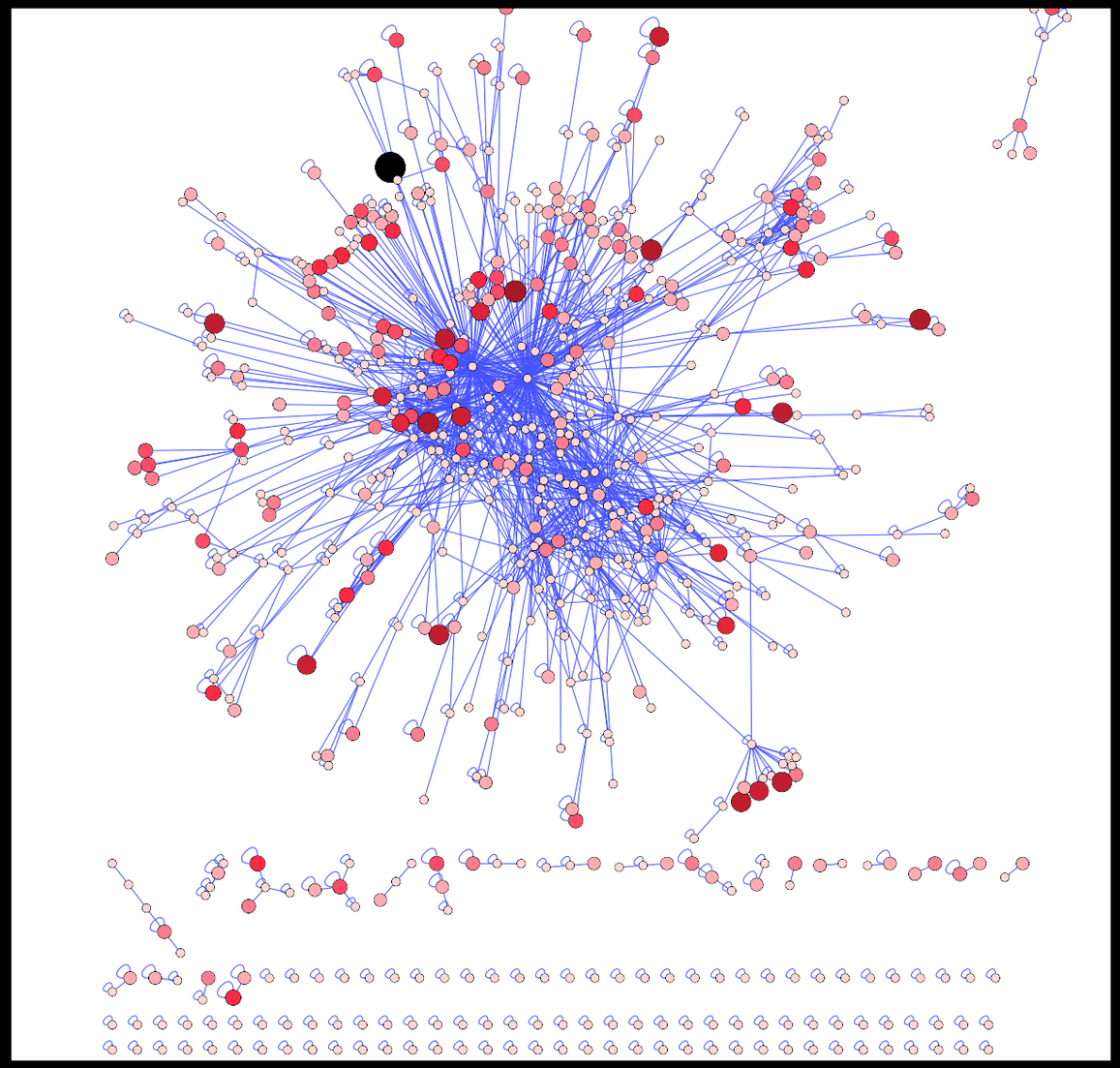


Various optimization functions & algorithms have been developed depending on the task.

Visualization & Interactive Exploration



Yeast protein interaction network



Clustered (via graph compression) of yeast interaction network

Graph Summarization

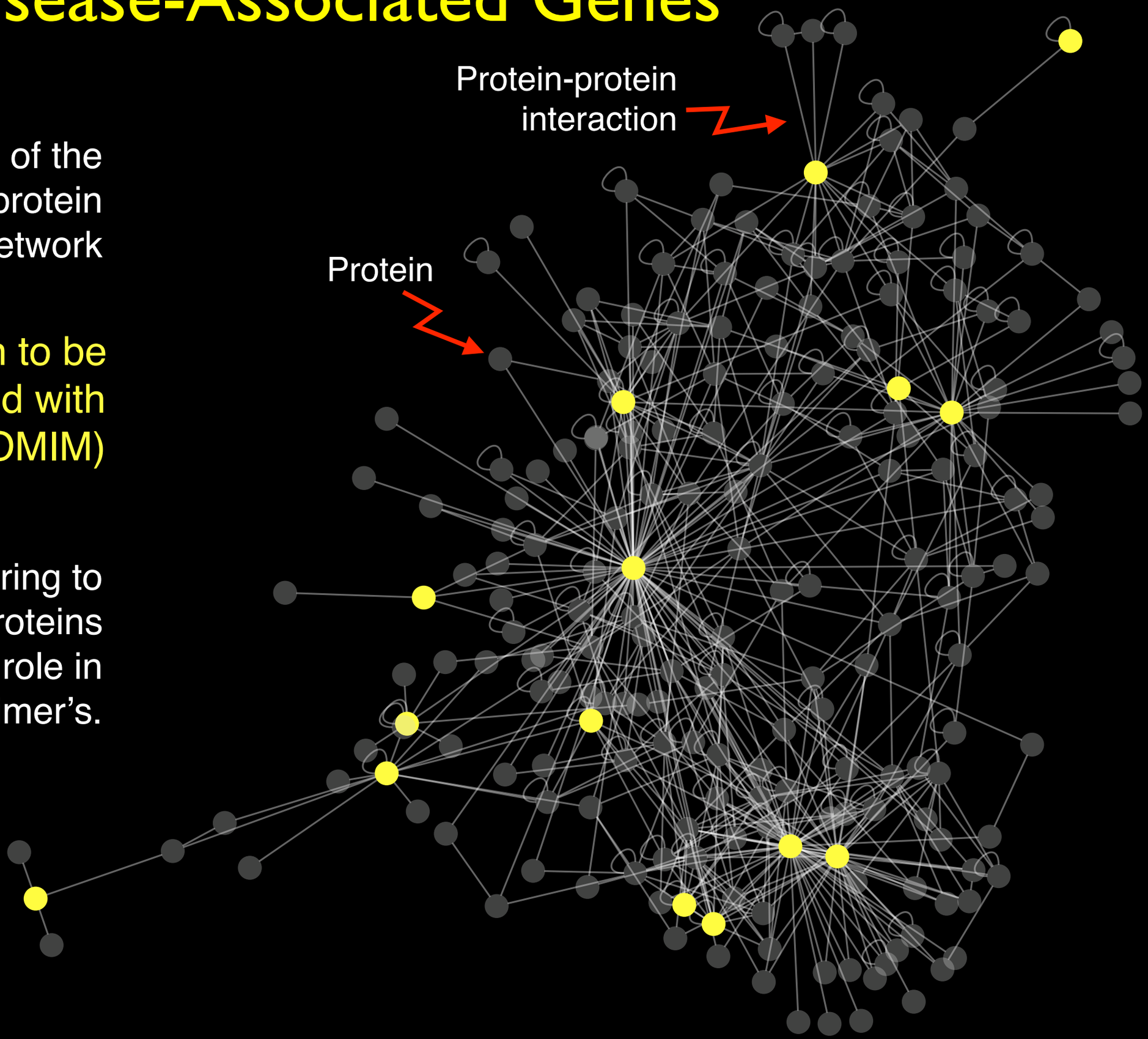
(Navlakha, Schatz, & K., J. Comp. Biol., 2009)

Finding Disease-Associated Genes

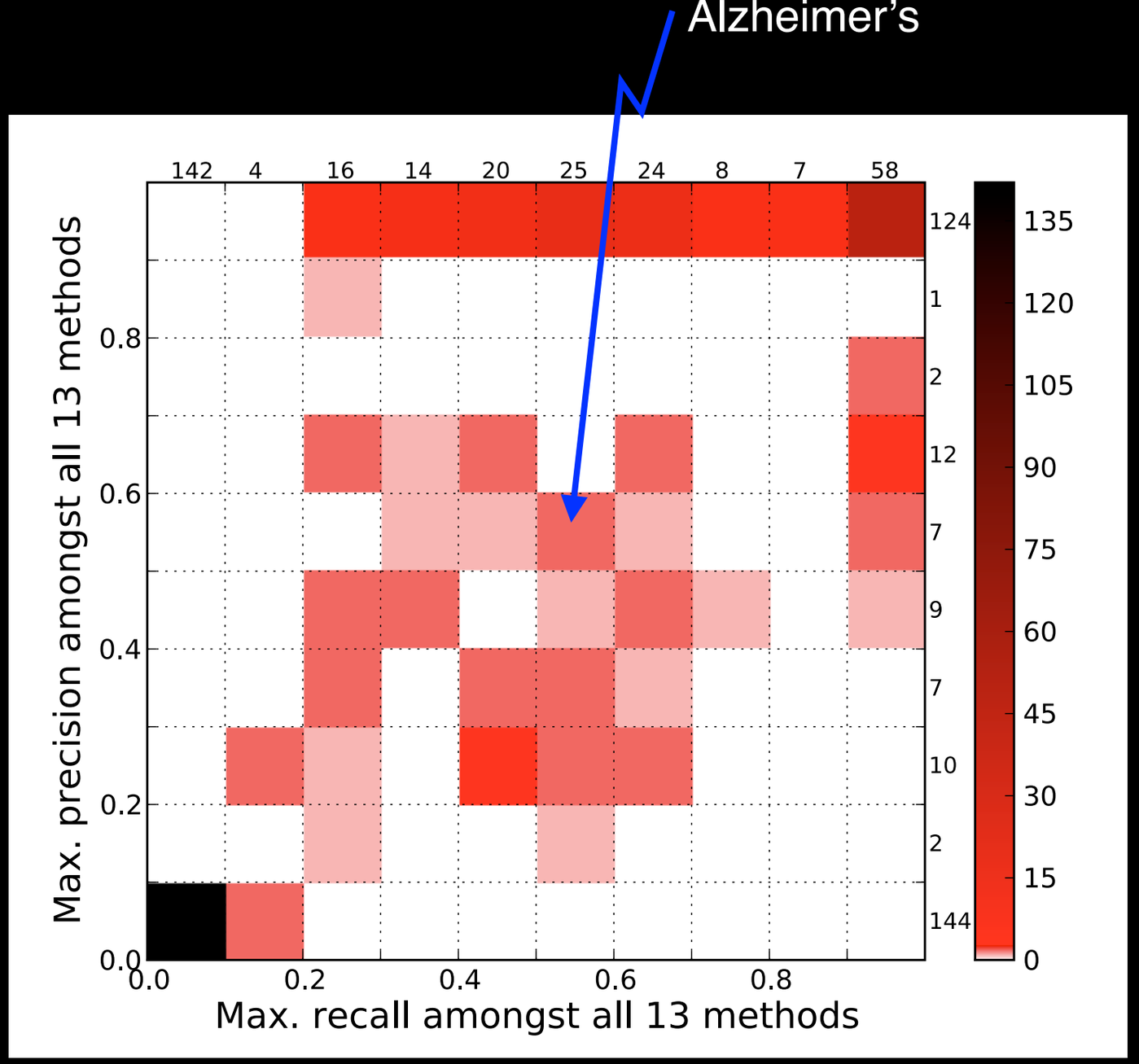
Small portion of the human protein-protein interaction network

Proteins known to be associated with Alzheimer's (OMIM)

Can use clustering to find additional proteins that play a role in Alzheimer's.



Finding Disease-Associated Genes



% of predictions that are correct.

% of omitted annotations recovered.

Function Transfer Methods:

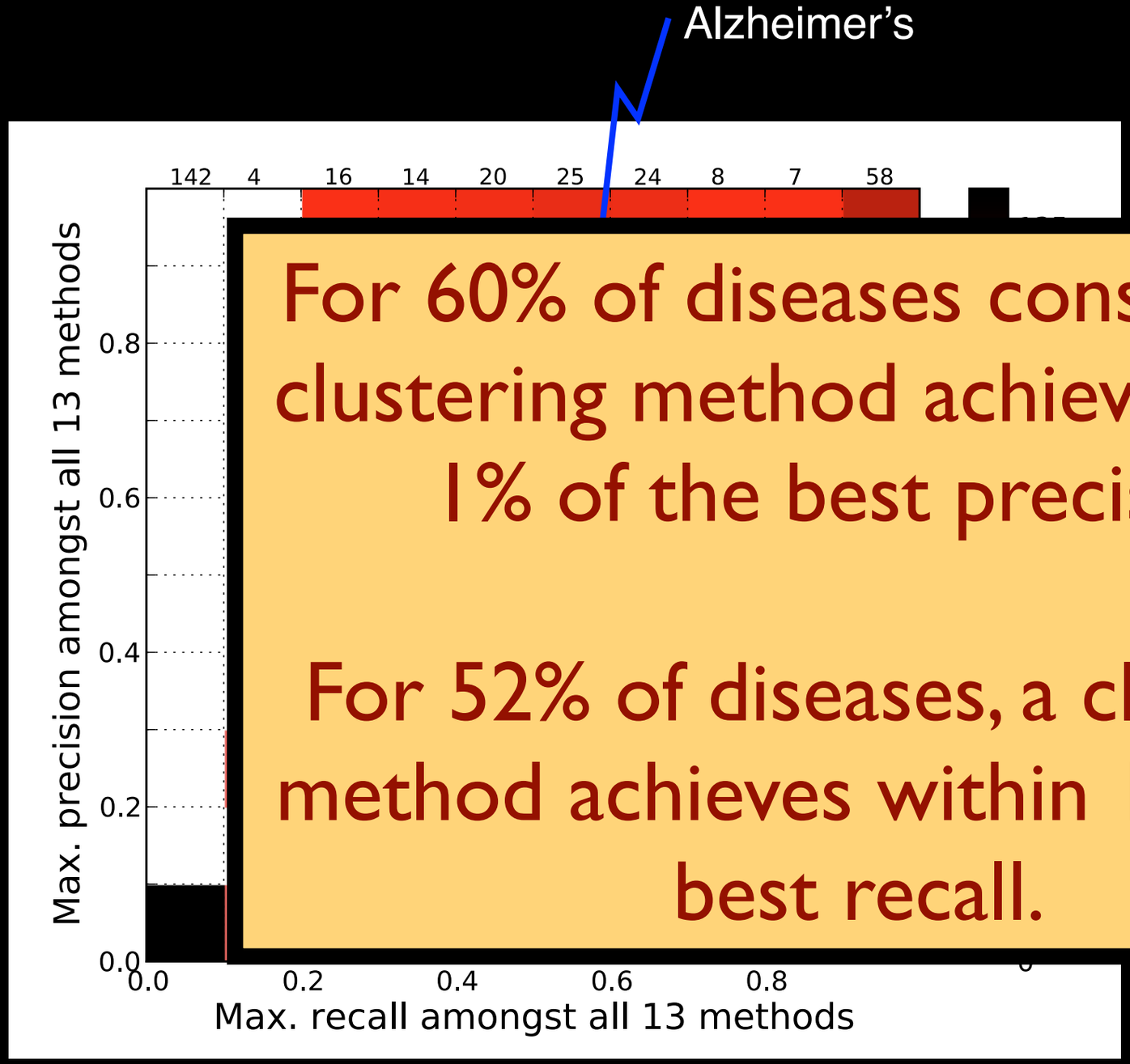
- Neighborhood (majority rule)
- Clustering (MCL, GS, VI-CUT)
- Random-walk (and flow-like algorithms)

Leave-one-out cross validation.

124 (out of 450) disease families with

(Navlakha & K., Bioinformatics, 2010)

Finding Disease-Associated Genes



For 60% of diseases considered, a clustering method achieves within 1% of the best precision.

For 52% of diseases, a clustering method achieves within 1% of the best recall.

% of predictions that are correct.

% of omitted annotations recovered.

Function Transfer Methods:

rhood
rule)
g (MCL,
CUT)
-walk (and
algorithms)

out cross

124 (out of 450) disease families with

(Navlakha & K., Bioinformatics, 2010)

Challenges With Clustering Biological Networks

1. Uncertainty in the clustering objective function & the quality of the optimal solution.

   Explore alternative clusterings in a systematic way.

2. Uncertain edge links due to experimental noise and limitations.

   Cluster probabilistic graphs.

3. Known – but possibly wrong – cluster membership for some nodes.

   Navlakha, White, Nagarajan, Pop, and Kingsford, RECOMB 2009.

Challenges With Clustering Biological Networks

1. Uncertainty in the clustering objective function & the quality of the optimal solution.

   Explore alternative clusterings in a systematic way.

2. Uncertain edge links due to experimental noise and limitations.

   Cluster probabilistic graphs.

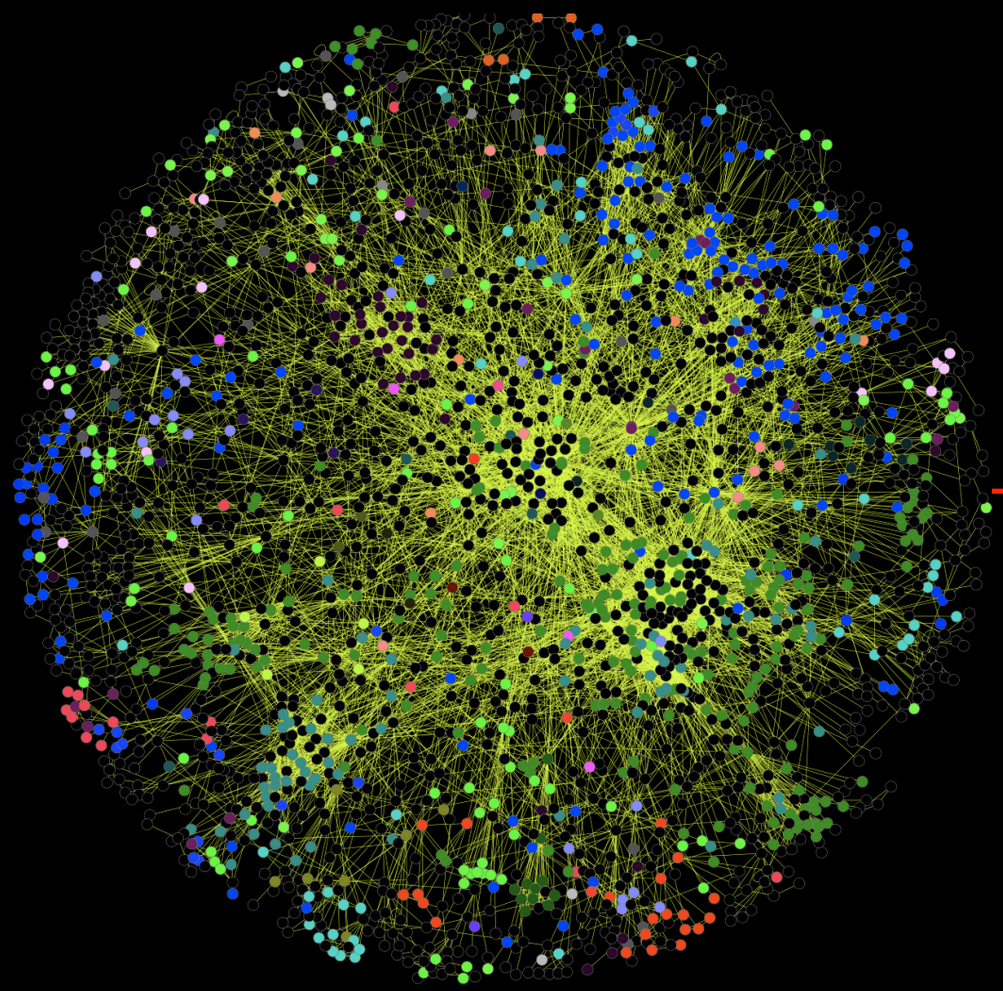
3. Known – but possibly wrong – cluster membership for some nodes.

   Navlakha, White, Nagarajan, Pop, and Kingsford, RECOMB 2009.

What can we learn from near-optimal clusterings?

Exploiting the degeneracy of modularity

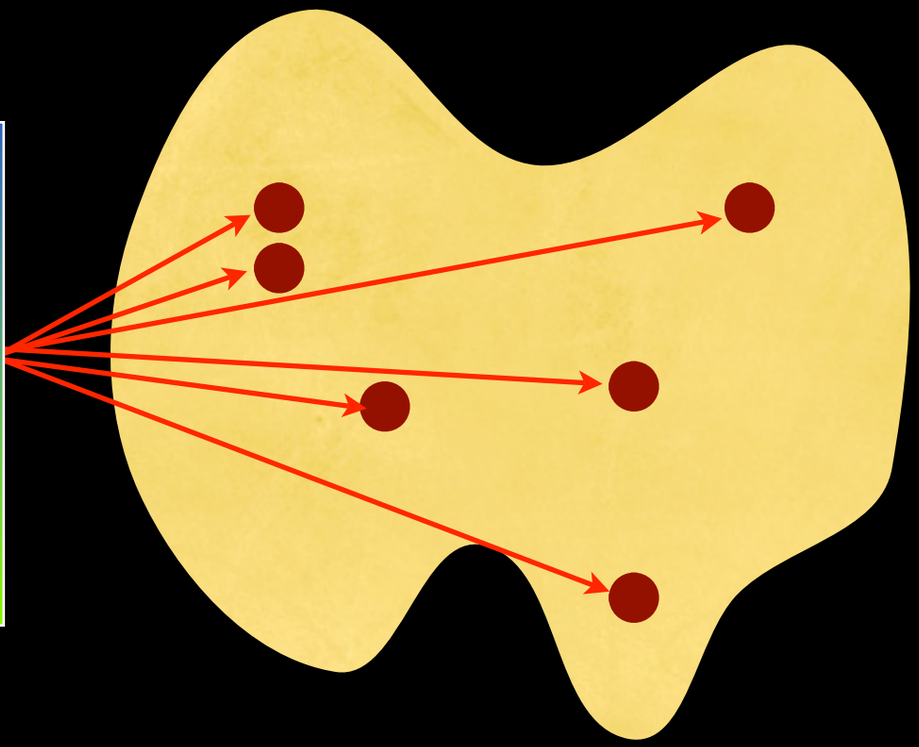
There Are Often Many Reasonable Network Clusterings



e.g. protein-protein interaction network

van Dongen 1998
Newman 2003
Bader and Hogue 2003
King+ 2004
Pereira-Leal+ 2004
Pons and Latapy 2005
Blatt+ 2006
Royer+ 2008
Navlakha+ 2009

clustering algorithm

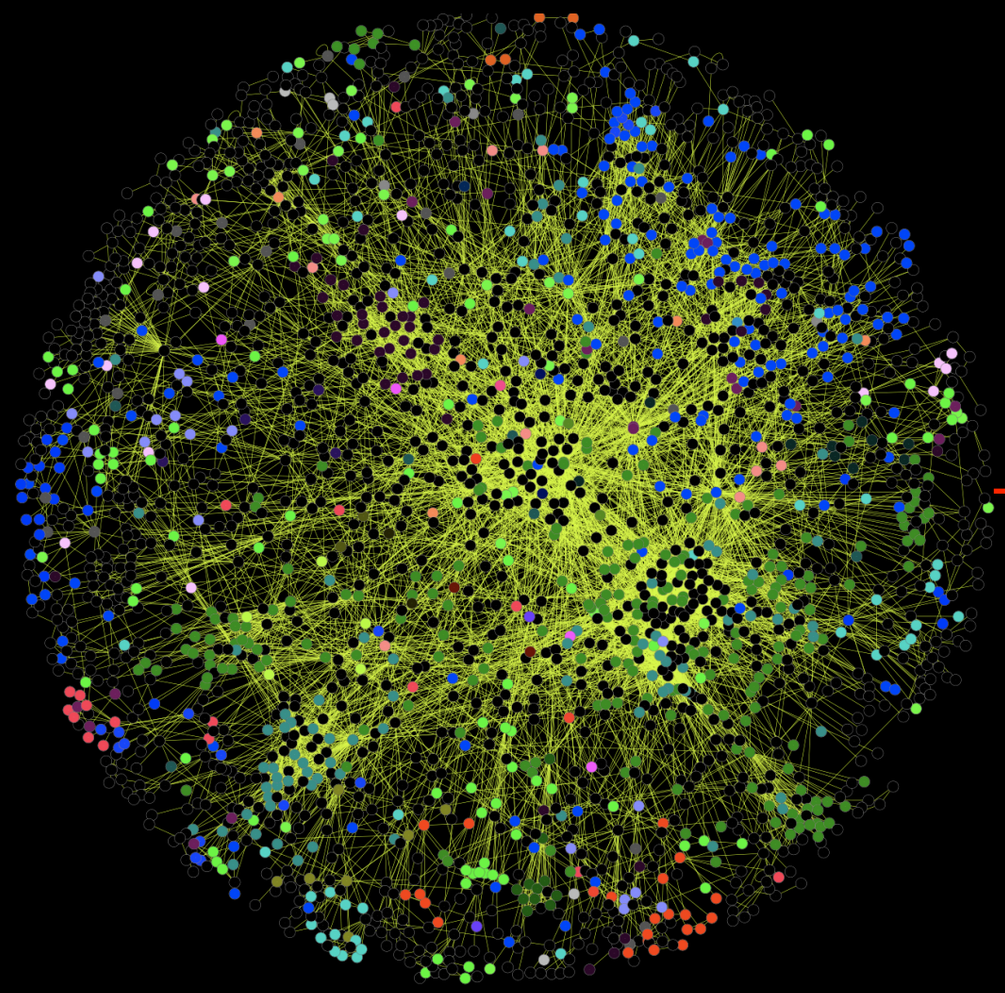


space of clusterings

How can we sample the space of network clusterings?

What can near-optimal clusterings reveal about network structure?

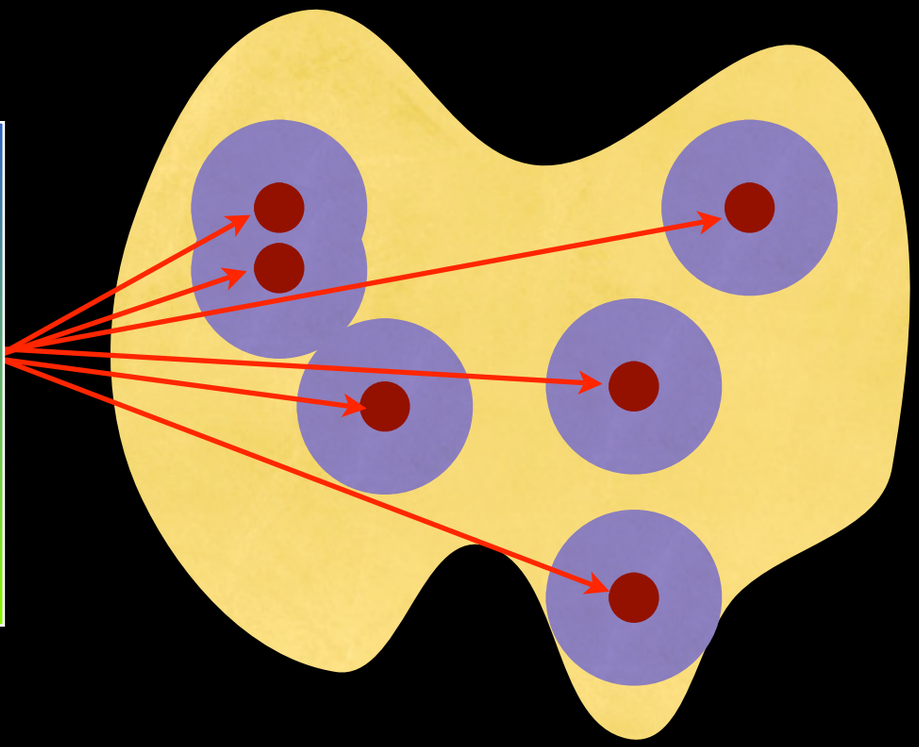
There Are Often Many Reasonable Network Clusterings



e.g. protein-protein interaction network

van Dongen 1998
Newman 2003
Bader and Hogue 2003
King+ 2004
Pereira-Leal+ 2004
Pons and Latapy 2005
Blatt+ 2006
Royer+ 2008
Navlakha+ 2009

clustering algorithm



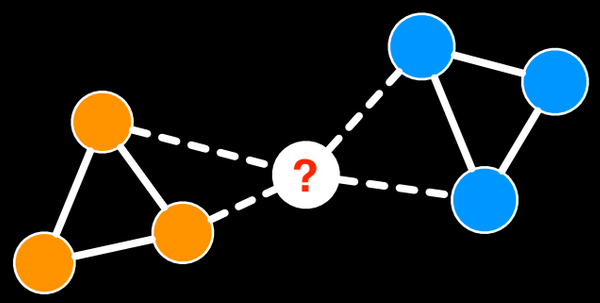
space of clusterings

How can we sample the space of network clusterings?

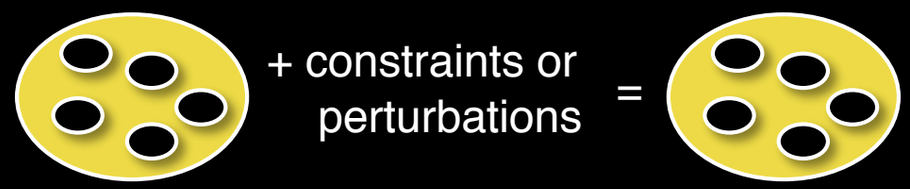
What can near-optimal clusterings reveal about network structure?

Usefulness of Near-Optimal Clusterings

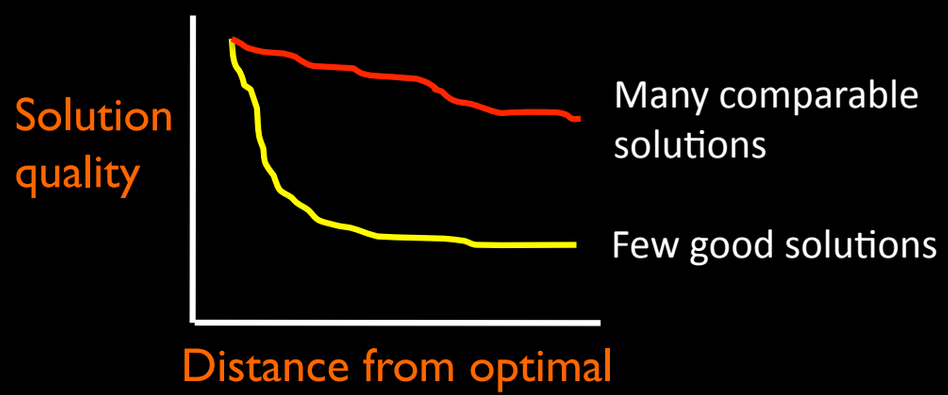
1. Correct clustering could be obscured by noise



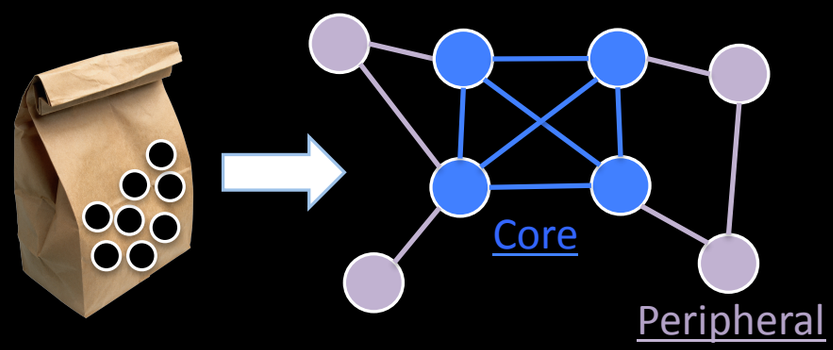
2. Resilient or robust communities can be identified



2. Confidence in the optimal solution can be assessed by comparing nearby solutions



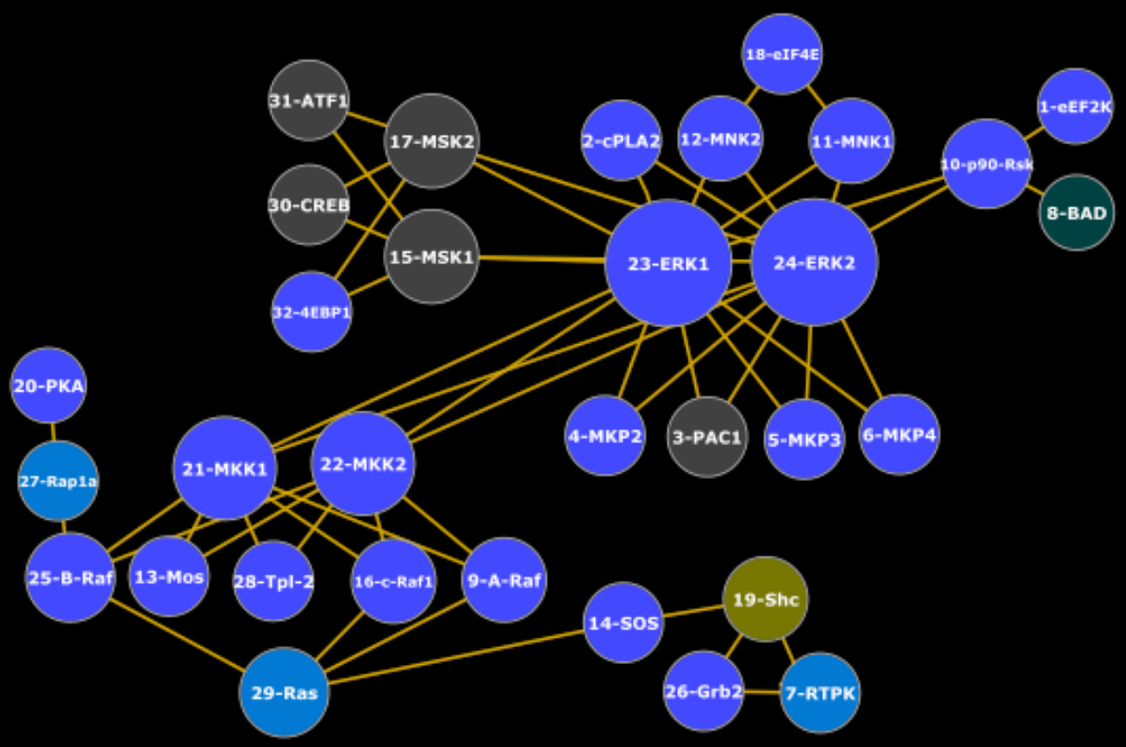
4. Core and peripheral community members can be distinguished

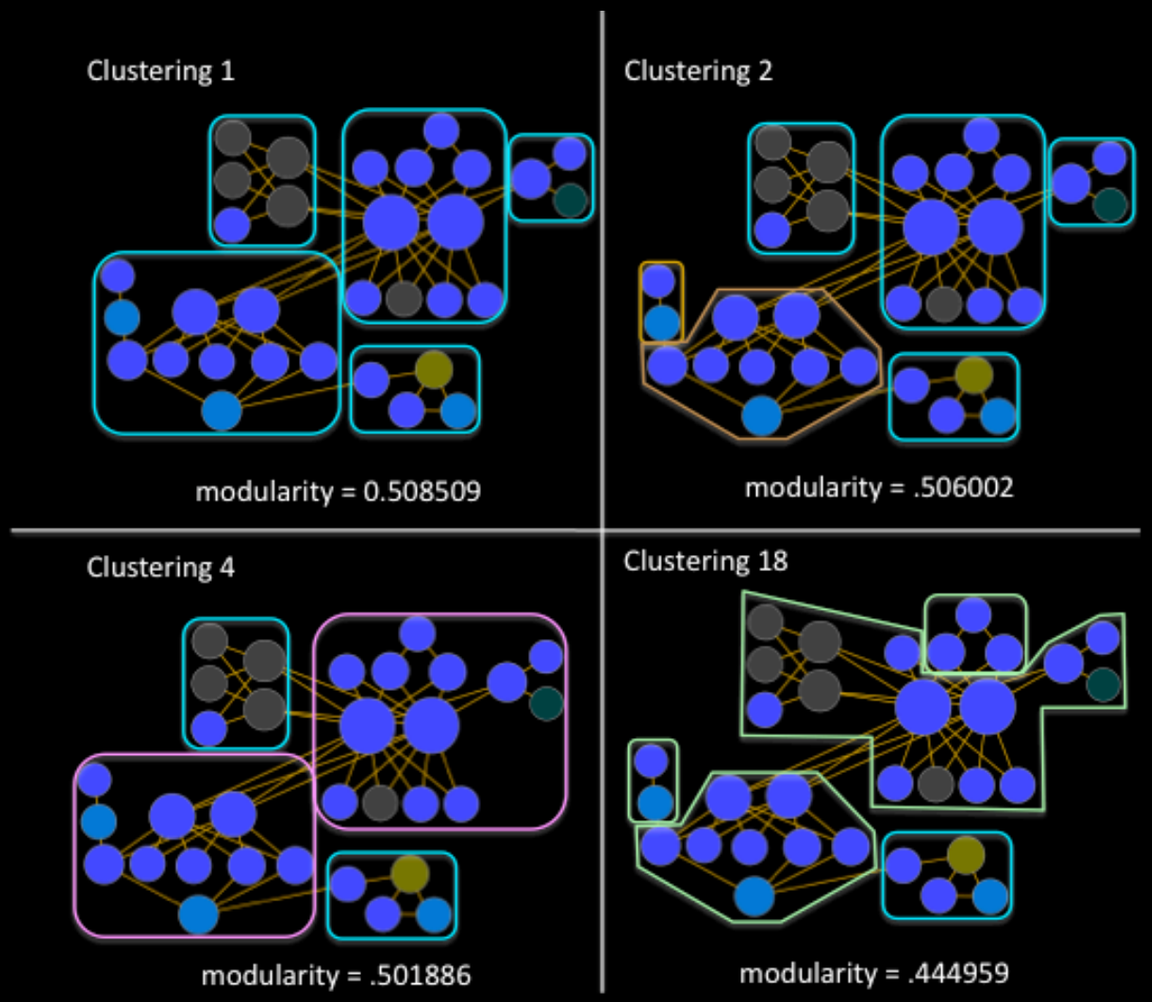


Near-Optimal Solutions Represent Legitimate Clusterings

Cluster quality: modularity (Girvan & Newman, 2003)

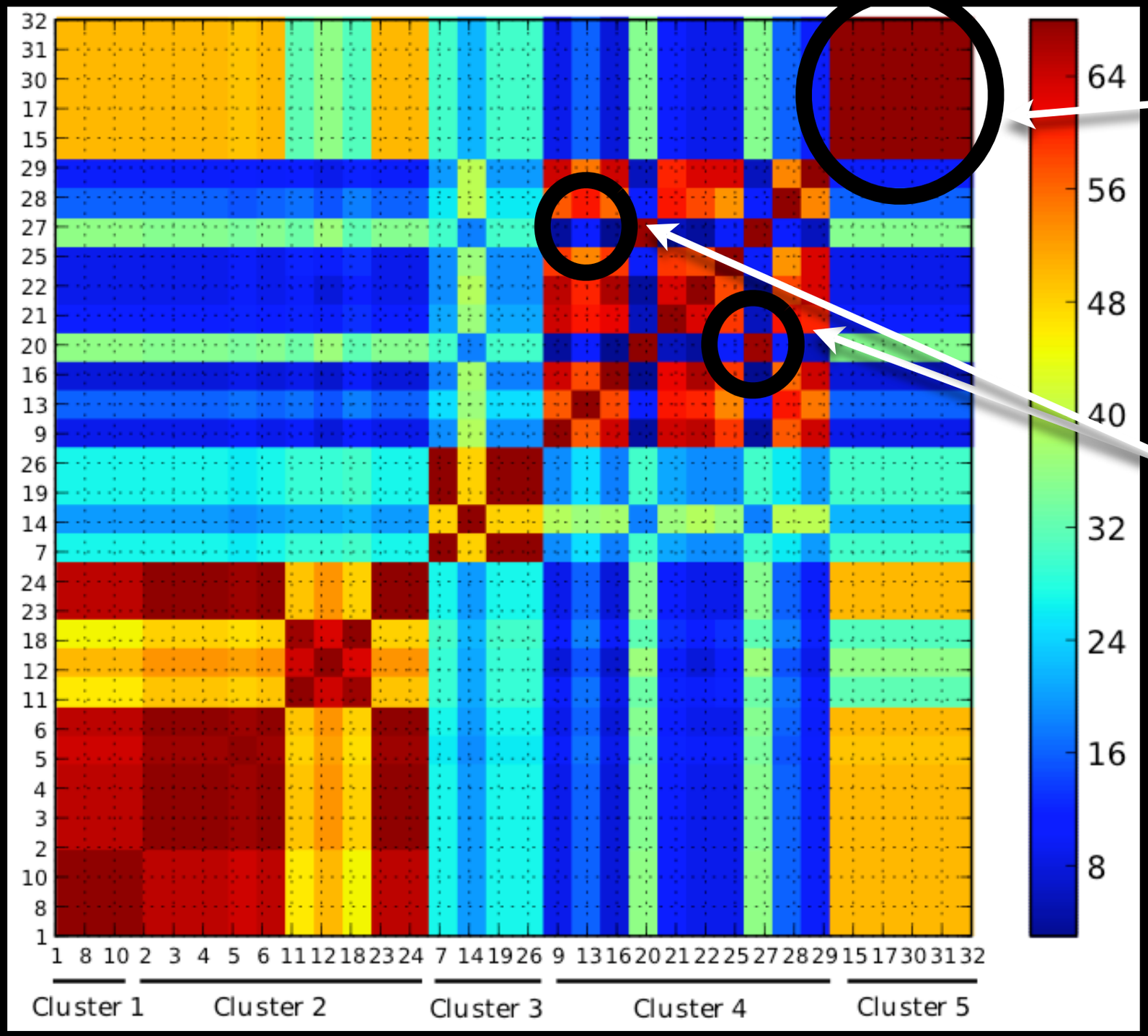
ERK1/ERK2 MAPK signaling pathway (32 nodes; 54 edges)





Highly conserved across eukaryotes; regulates meiosis, mitosis, metabolism

MAPK Global Clustering Dynamics



Co-clustering heatmap

Resilient community

Nodes remain together in many near-optimal solutions

Core & peripheral members

Nodes 27 (PKA) & 20 (Rap1a) travel together more than 27 and 13 (Mos)

Any single solution would miss these dynamics

How can we generate many good near-optimal clusterings?

- Simulated annealing (e.g. Guimera+ 2005, Massen & Doye 2005): global optimization heuristic

- Linear programming (Agarwal & Kempe 2008): randomized rounding to convert fractional solutions to integral solution

- Random perturbation (e.g. Hadjitodorov+ 2006, Nabieva+ 2005, Hopcroft+ 2004): cluster once, randomly perturb objective function or data, re-cluster.

- All based on randomness:
  - Large deviations are improbable
  - No guarantee that perturbed solutions are of high-quality

⇒ To avoid these problems, we propose a constraint-based approach.

Modularity

A clustering quality measure

Intuition: want # of edges inside a cluster > expected by chance and # of edges between clusters < expected by chance.

$$q(G, \mathcal{C}) := \sum_{u,v \in V} \left(A_{uv} - \frac{k_u k_v}{2m} \right) (1 - x_{uv})$$

Adjacency matrix of G

Degree of node *v*

edges in G

0 if u and v are in the same cluster; 1 otherwise

Integer-Linear Program (ILP) to Maximize Modularity (Agarwal+, 2008)

$$\text{maximize} \quad \sum_{u \in V} \sum_{v \in V} (A_{uv} - \frac{k_u k_v}{2m})(1 - x_{uv})$$

$$\text{subject to}$$

$$x_{uv} + x_{vw} \geq x_{uw} \qquad \text{for all } u, v, w \in V$$

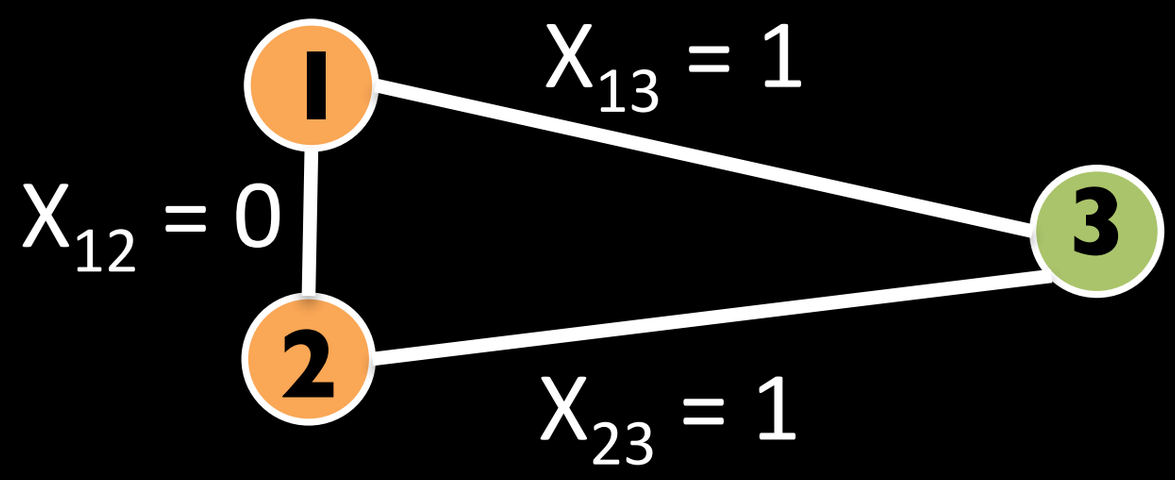
Enforce triangle inequality to make a legitimate partition

$$x_{uv} \in \{0, 1\}$$

Hard clustering (not soft)

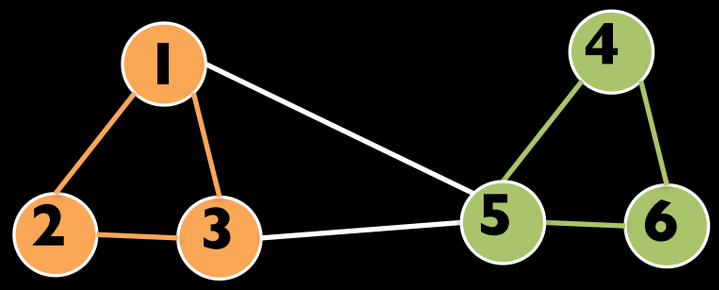
Solving this ILP is NP-hard but for reasonably small networks it can be solved optimally using branch-and-bound.

Can use heuristic solutions as initial LP basis and lower bounds.



Solution Vector

Input network



Optimal modularity clustering

x^0 = [– 0 0 1 1 1

– – 0 1 1 1 ← nodes 3 and 6 are not co-clustered

– – – 1 1 1

– – – – 0 0 ← nodes 5 and 6 are co-clustered

– – – – – 0]

= <0 0 1 1 1 0 1 1 1 1 1 1 0 0 0>

0	0	1	1	1	0	1	1	1	1	1	1	0	0	0

A solution is a (n choose 2)-sized vector

Diversity Constraints

pairs co-clustered in solution X^0 but not in X:

$$(\vec{1} - X^0) \cdot X \geq d^0_{\text{split}}$$

pairs in different clusters in X^0 but co-clustered in X:

$$X^0 \cdot (\vec{1} - X) \geq d^0_{\text{merge}}$$

Add combined constraint to ILP and re-solve:

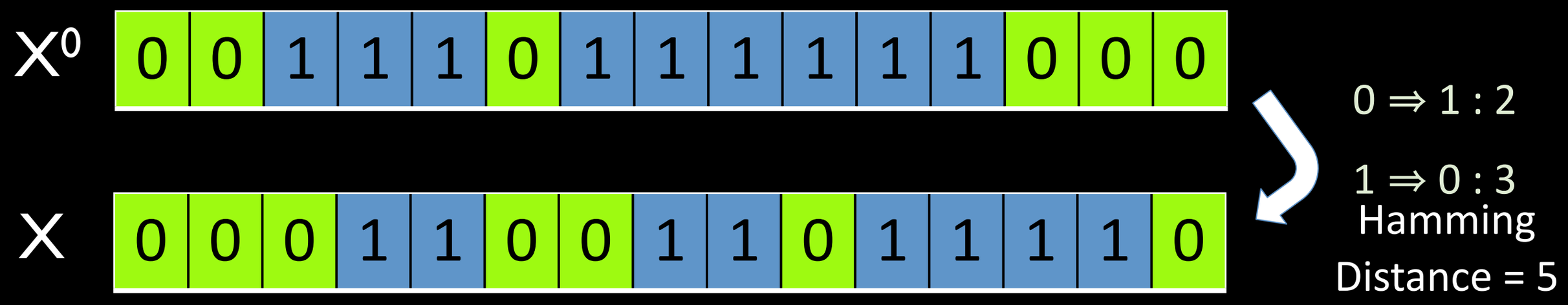
$$\underbrace{X^0 \cdot (\vec{1} - X) + (\vec{1} - X^0) \cdot X}_{\text{Hamming distance } \Delta(X,\, X^0) \text{ between X and } X^0} \geq d^0_{\text{changes}}$$

Diversity Constraints

Measure distance between clusterings by the # changes in the co-clustering matrix (Rand-index-like).

For alternative ways to measure this distance, see Geet Duggal & Saket Navlakha's poster.

$$X^0 \cdot (\vec{1} - X) + (\vec{1} - X^0) \cdot X \geq d^0_{\text{changes}}$$



Can iteratively add such constraints to explore more and more different solutions.

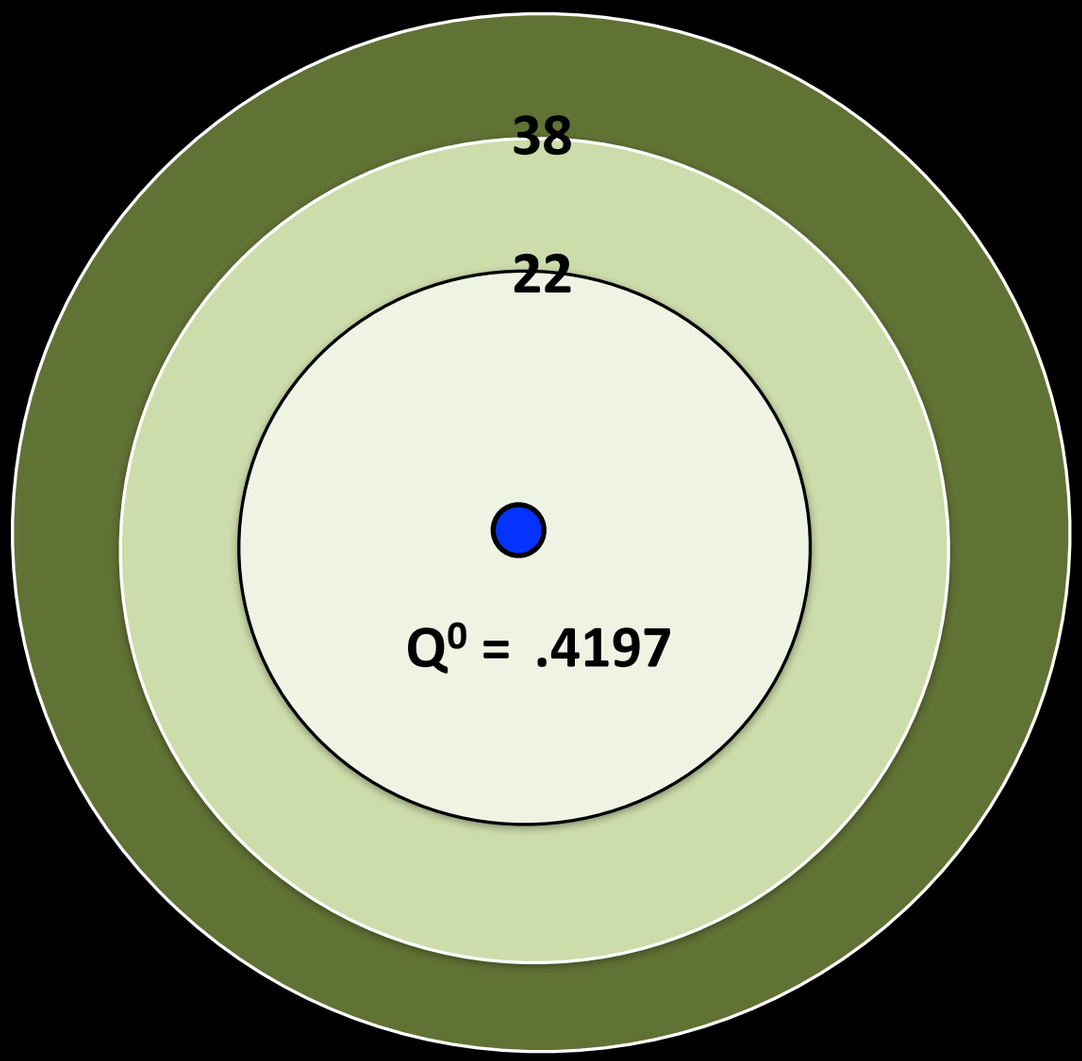
Creating the Modularity Landscape

Coarse grained:
quickly sample diverse solutions

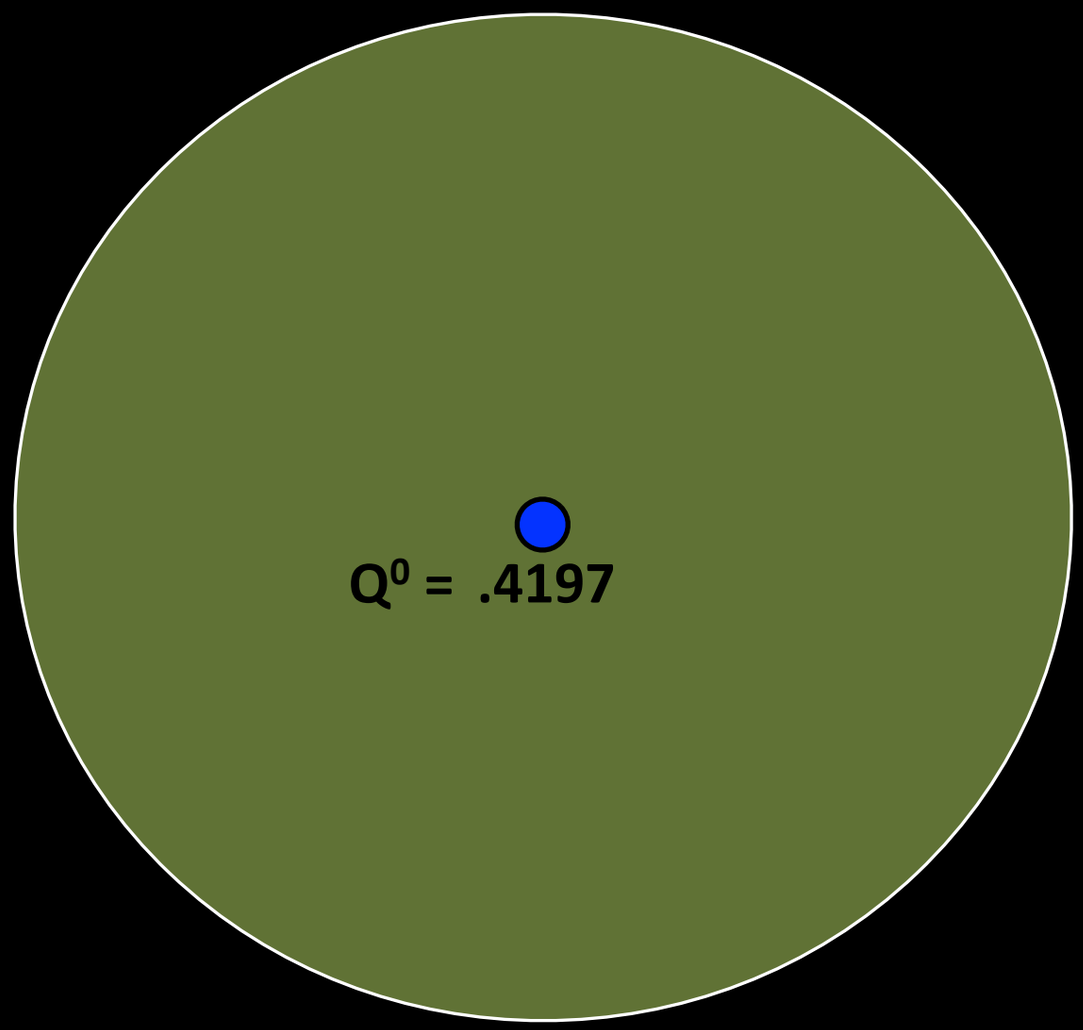
$$d_{changes}^{i+1} = \Delta(X^0, X^i) + 1$$

Fine grained:
i^{th} solution is a provably i^{th} optimal

$$\Delta(X^j, X^i) \geq 1 \; \forall j, 0 \leq j < i$$



Distance-based



Point-based

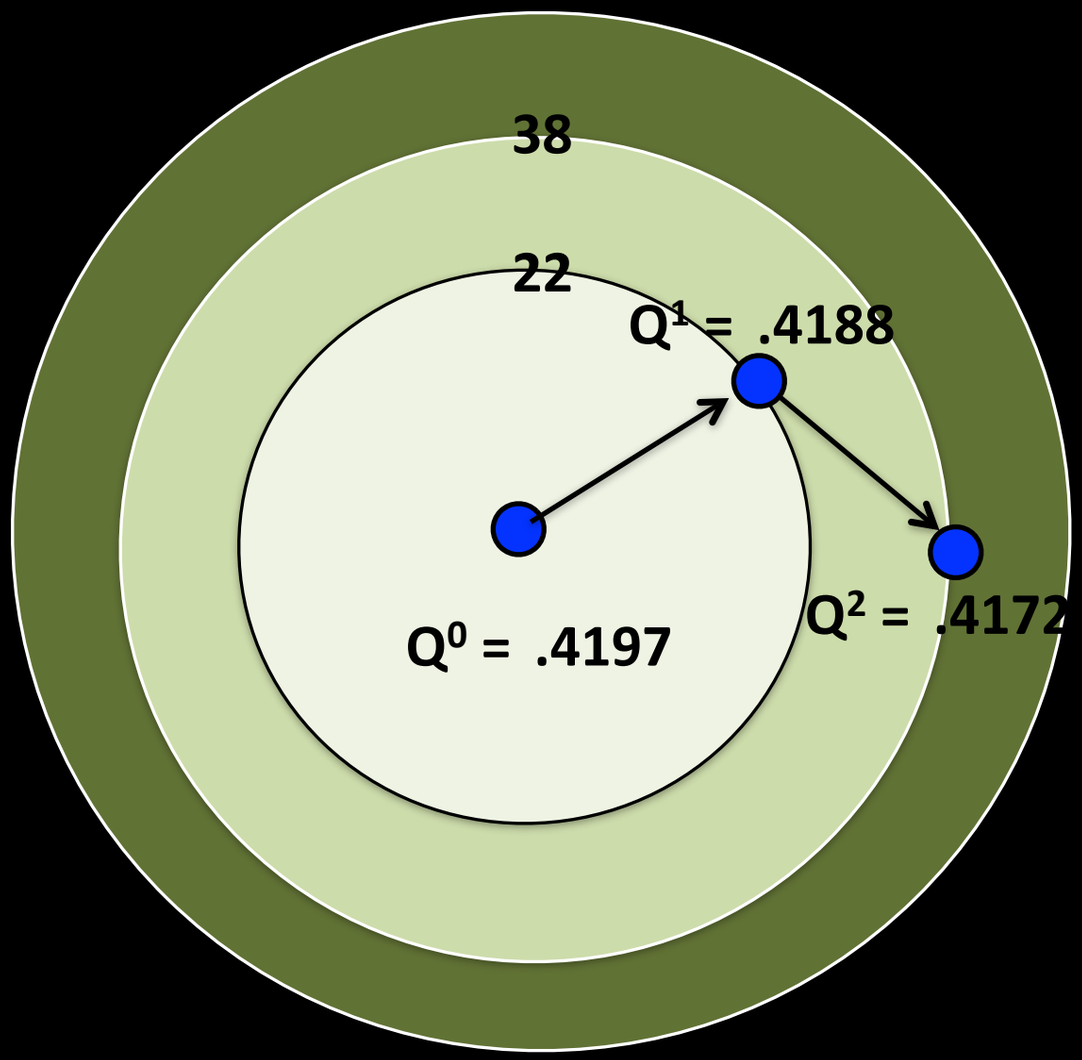
Creating the Modularity Landscape

Coarse grained:
quickly sample diverse solutions

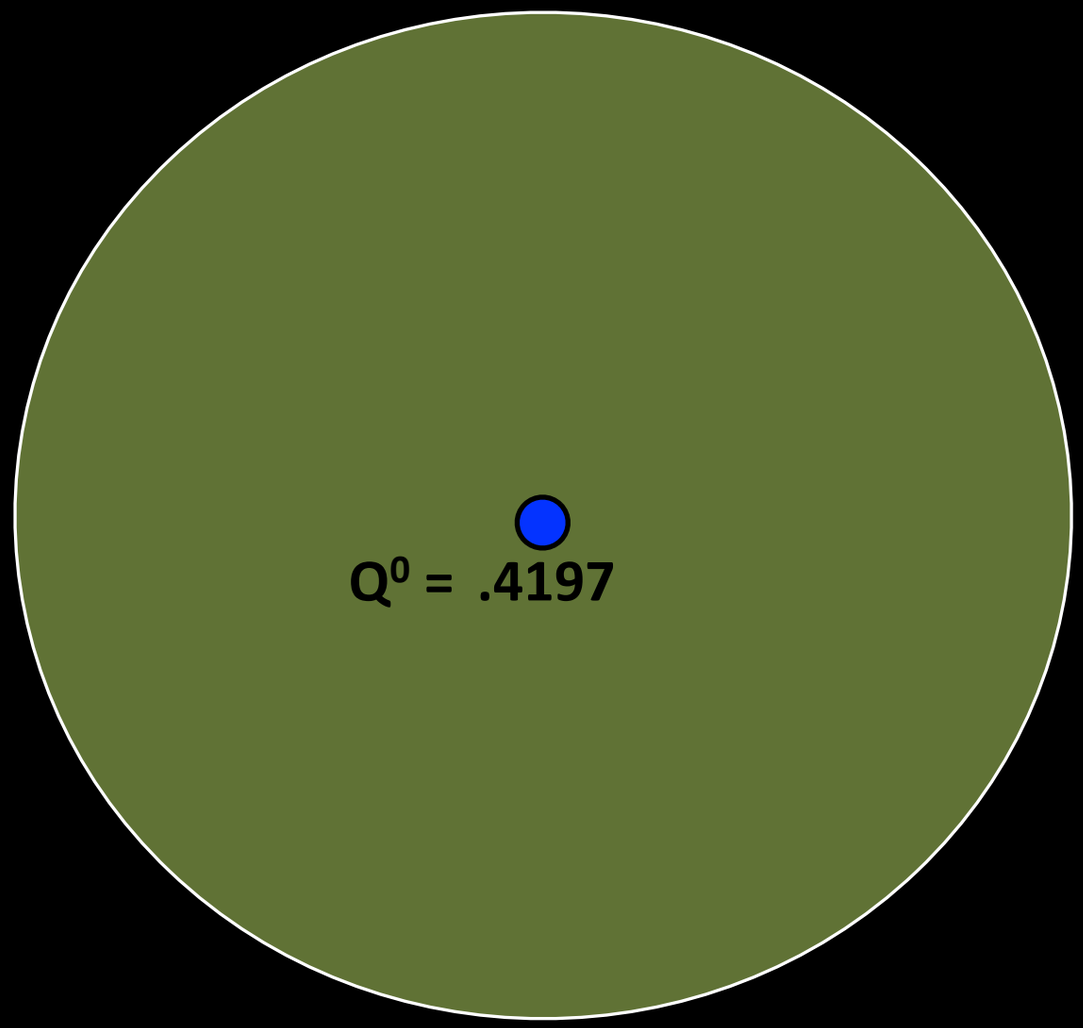
$$d_{changes}^{i+1} = \Delta(X^0, X^i) + 1$$

Fine grained:
i^{th} solution is a provably i^{th} optimal

$$\Delta(X^j, X^i) \geq 1 \ \forall j, 0 \leq j < i$$



Distance-based



Point-based

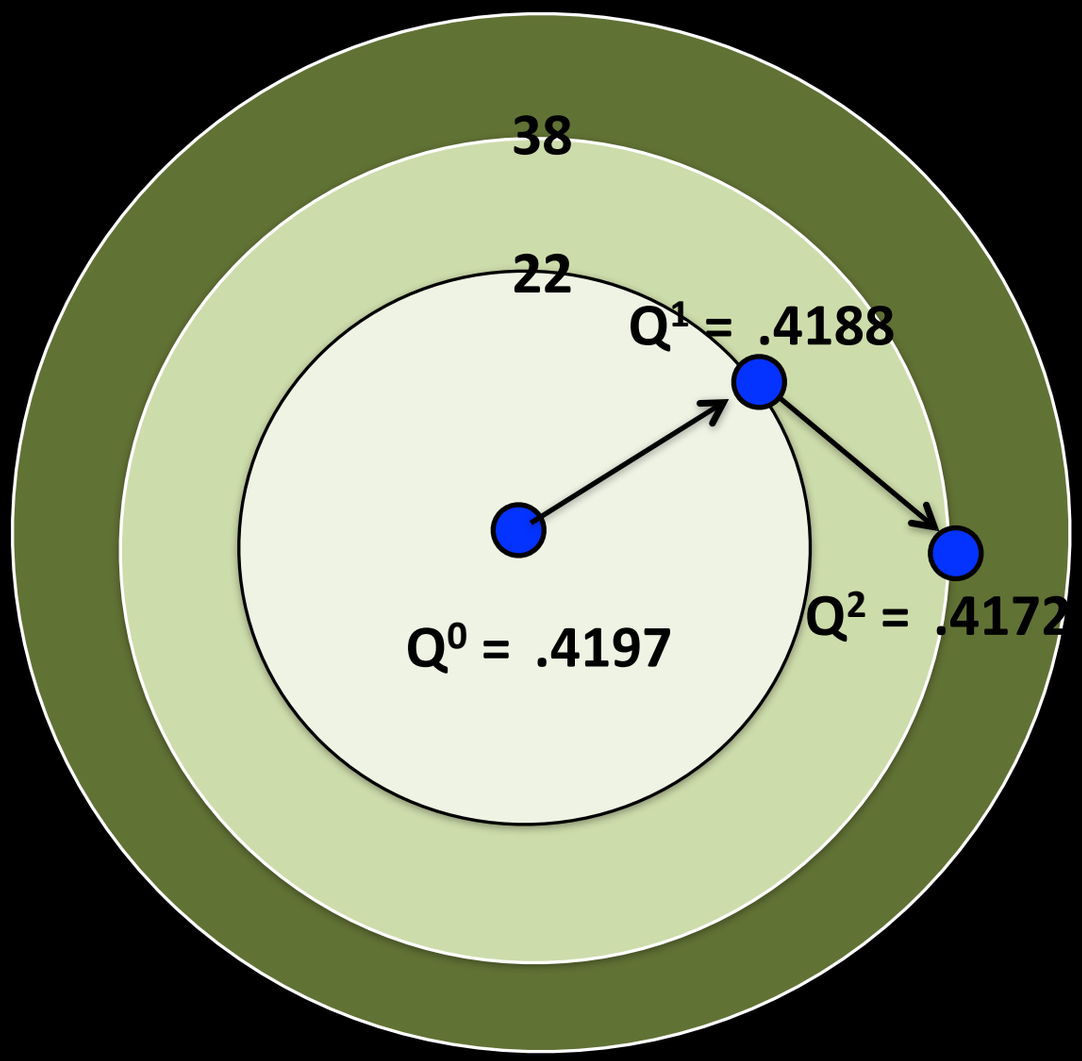
Creating the Modularity Landscape

Coarse grained:
quickly sample diverse solutions

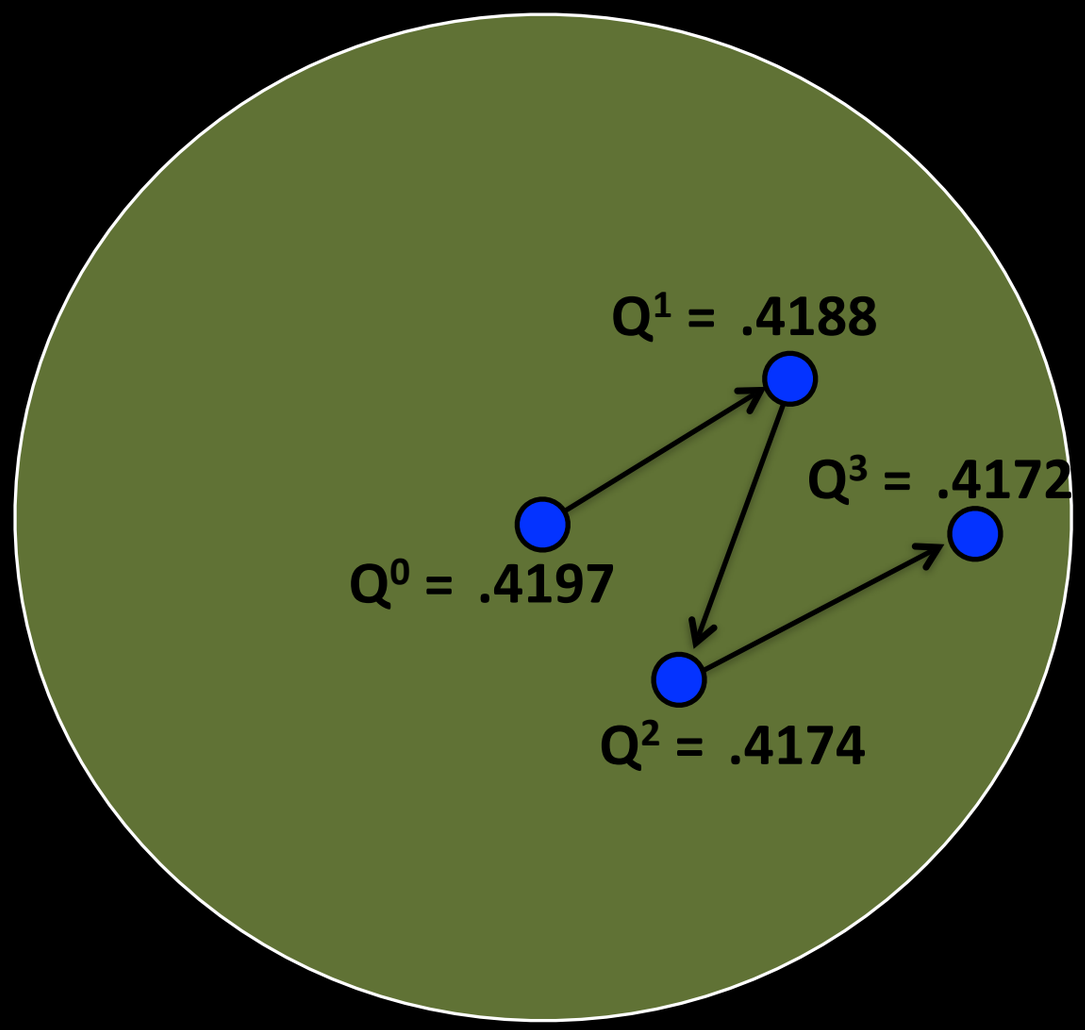
$$d_{changes}^{i+1} = \Delta(X^0, X^i) + 1$$

Fine grained:
i^{th} solution is a provably i^{th} optimal

$$\Delta(X^j, X^i) \geq 1 \; \forall j, 0 \leq j < i$$



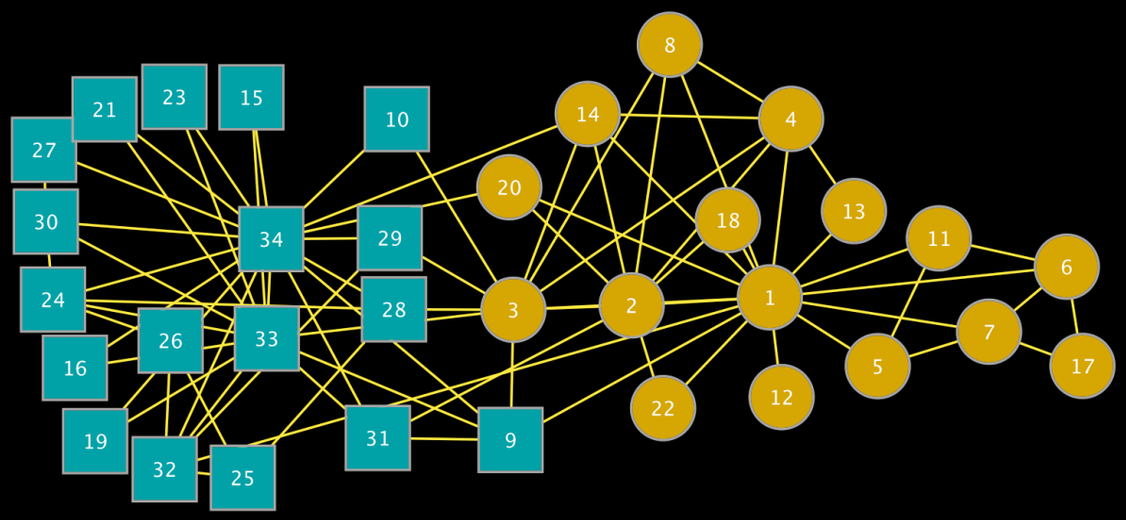
Distance-based



Point-based

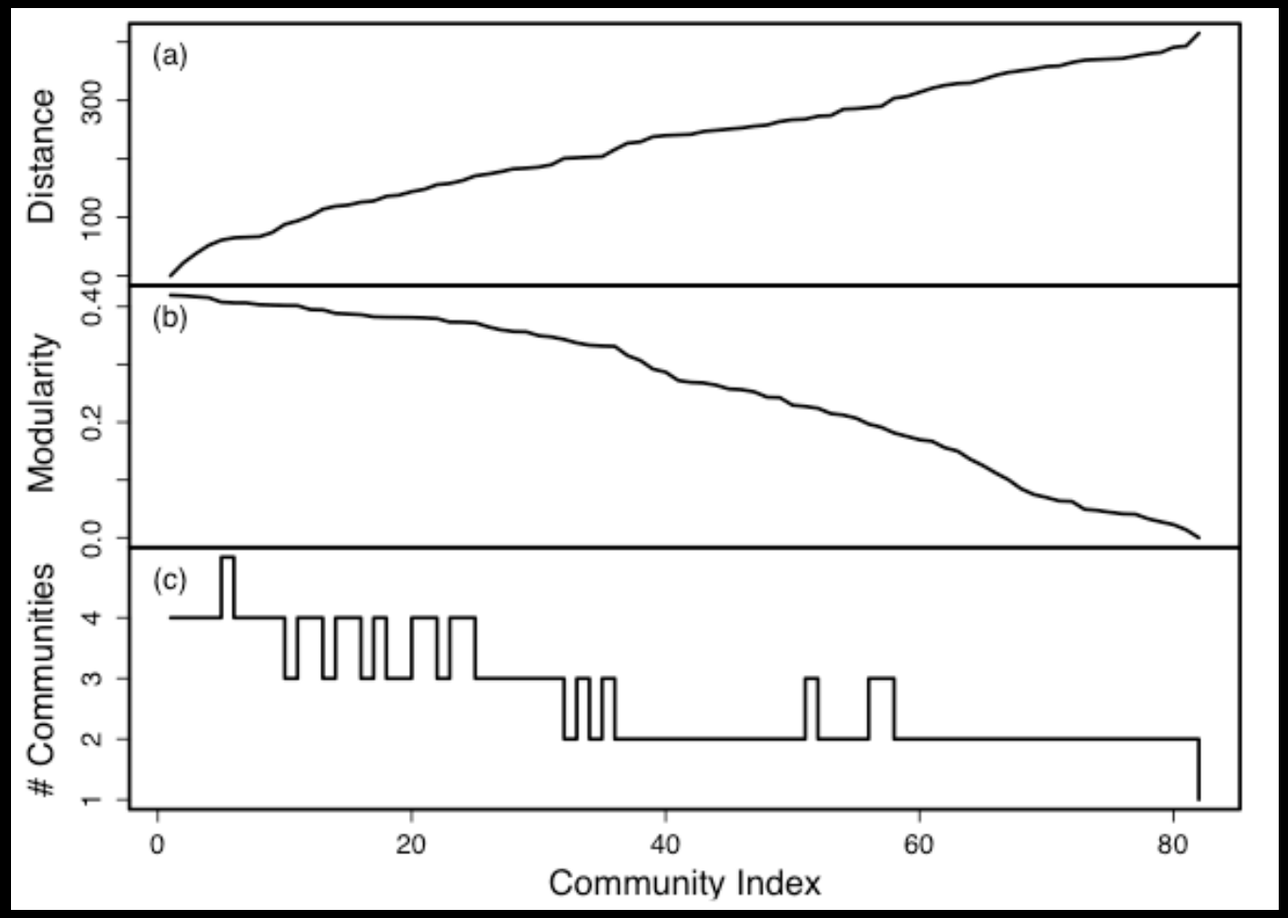
What can near-optimal solutions reveal about network structure?

Zachary's Karate Club (34 nodes, 78 edges)



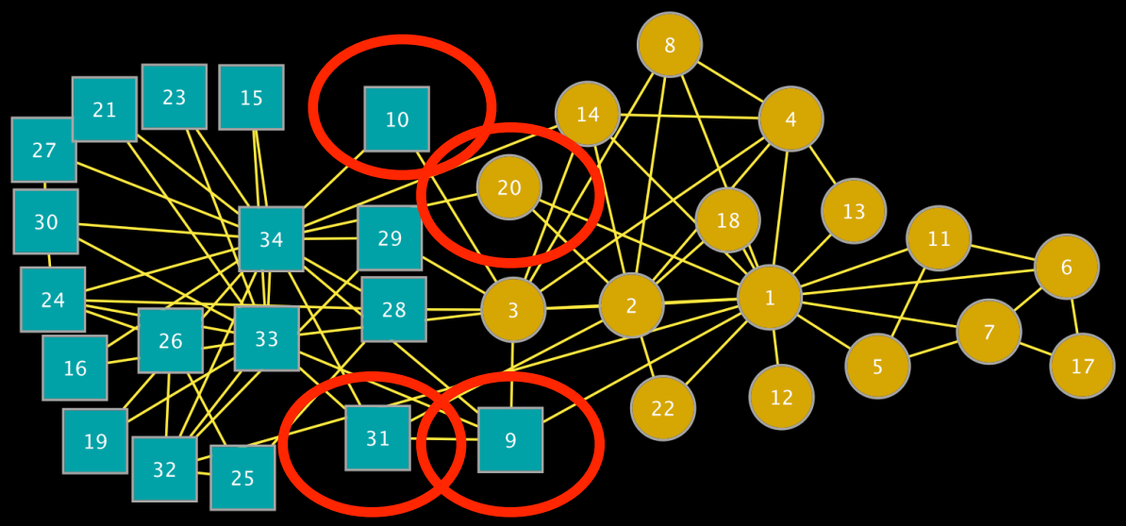
Linear relaxation returned an integral solution (so randomized rounding can't be done).

OPT mod = 0.419; 100th solution of **point-based** approach still has modularity > 0.4



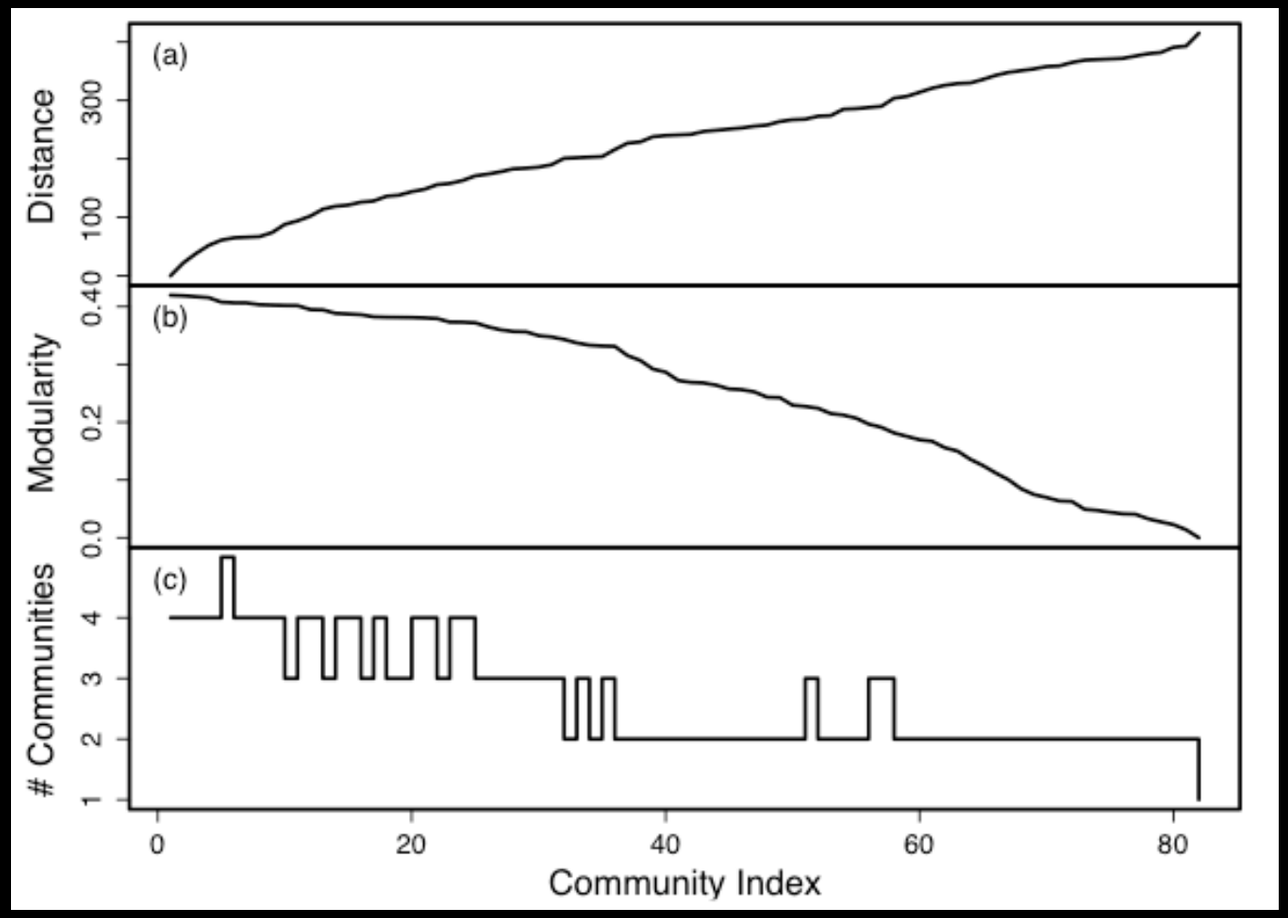
communities does not change monotonically (get to the "truth" of 2 communities at the 31st solution).

Zachary's Karate Club (34 nodes, 78 edges)



Linear relaxation returned an integral solution (so randomized rounding can't be done).

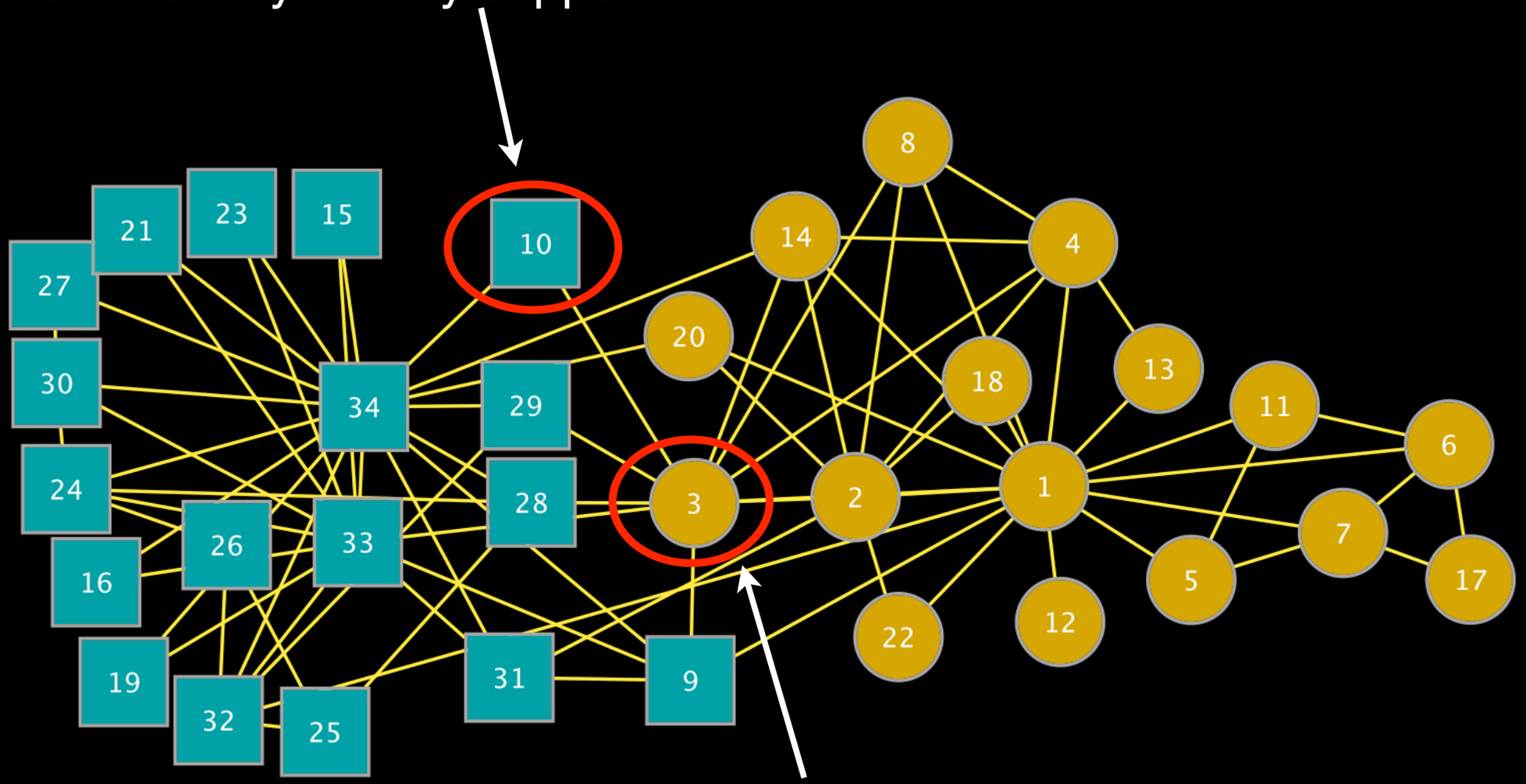
OPT mod = 0.419; 100th solution of **point-based** approach still has modularity > 0.4



communities does not change monotonically (get to the "truth" of 2 communities at the 31st solution).

Node 10 is with officer's faction in optimal, but instructor's group in 2nd best solution.

Known to only weakly support each faction.



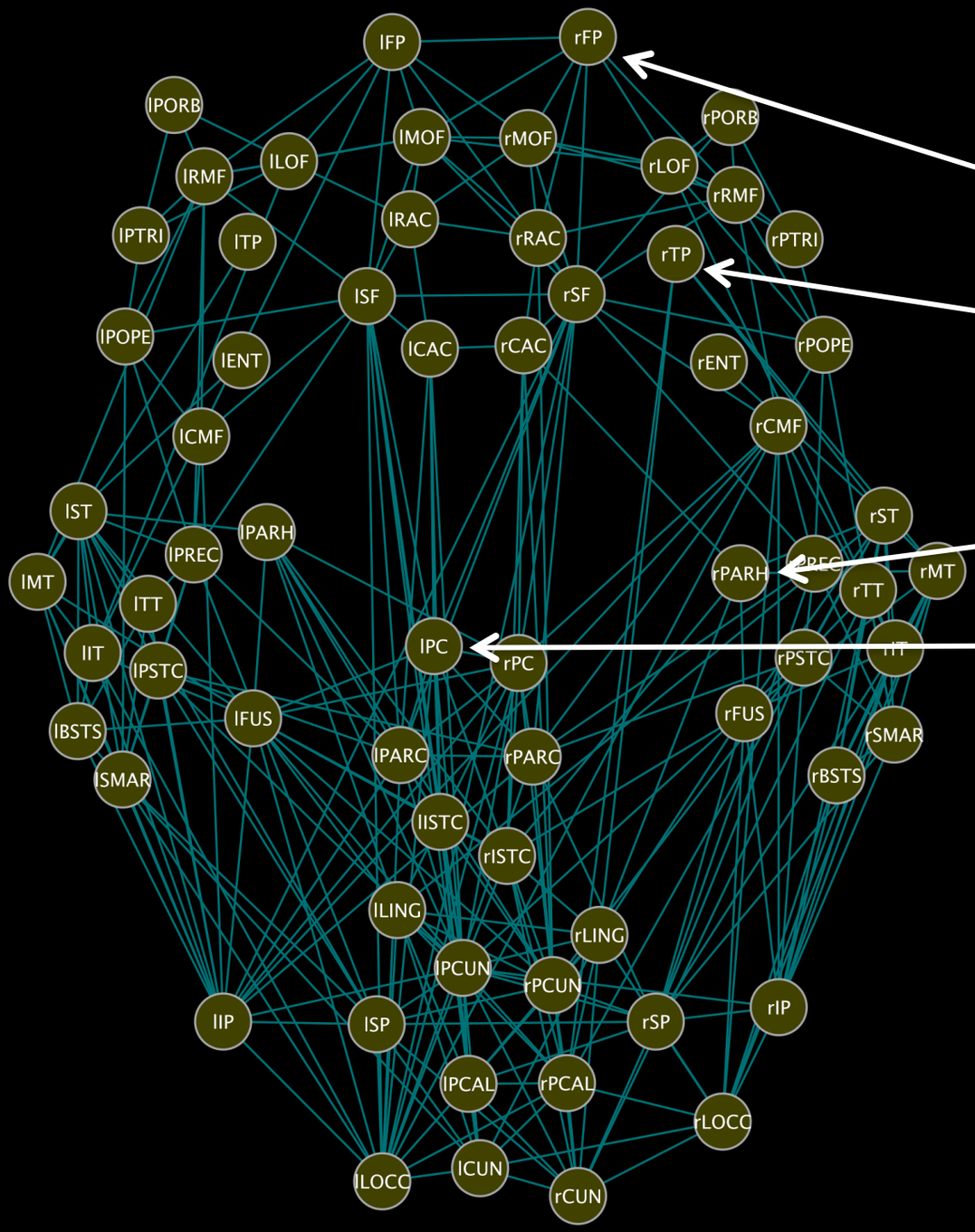
Equally topologically connected to both groups.

But known to be a strong supporter of the instructor.

Stays connected to instructor's group until the you get very far away from the optimal (@ the 72nd, 78th, and 80th near-optimal solutions)

Anatomical Brain Network (66 nodes, 2149 edges)

Diffusion spectrum imaging (DSI)



Hagmann+ 2008 and Brede Database

Fixed spatial positions:

Right frontal pole

Right temporal pole

Right parahippocampal cortex

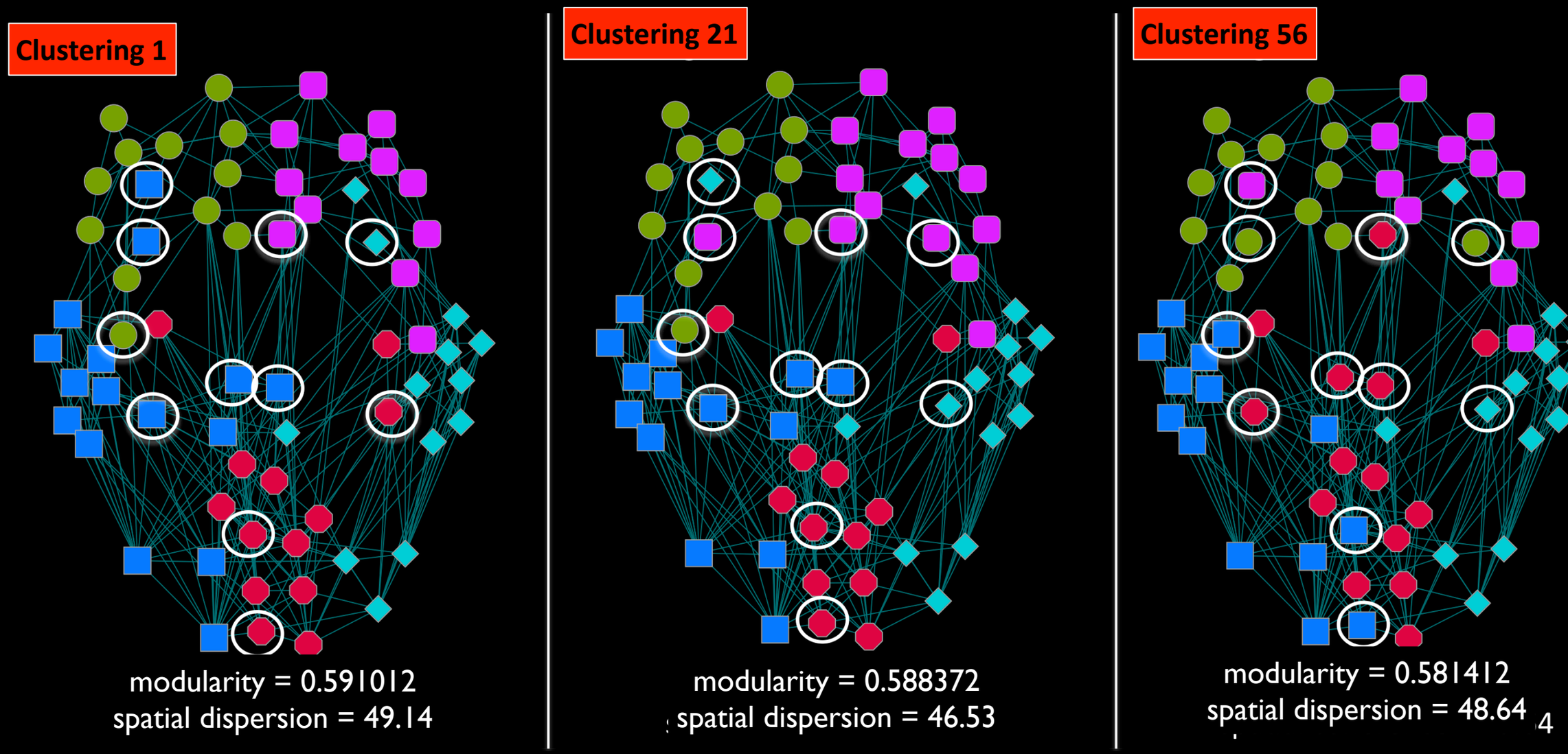
Left posterior cingulate cortex

Convert weighted edges to multiedges:

A_{uv} = # of edges between *u* and *v*

(instead of just 0 or 1)

Uncertainty in the optimal clustering



Degeneracy in modularity:

First 53 solutions (distance-based) all within 1% of optimal modularity

Near-optimal solutions also have better spatial cohesion

Near Optimal Solutions

- Near-optimal solutions provide deeper insight into community structure and dynamics:
  - Combat noise in network
  - Assess confidence in optimal clustering
  - Identify resilient communities
  - Distinguish between core and peripheral members

- Explicit constraint-based approach using integer linear programming
  - Theoretically sound with distance and optimality guarantees

- Future work
  - Near-optimal solutions for other objective functions?
  - Larger networks; core-peripheral proteins in complexes [Gavin+ 2006]
  - Do dynamics across the landscape correlate with dynamics over time?
  - Improve running time

Challenges With Clustering Biological Networks

1. Uncertainty in the clustering objective function & the quality of the optimal solution.

   Explore alternative clusterings in a systematic way.

2. Uncertain edge links due to experimental noise and limitations.

   Cluster probabilistic graphs.

3. Known – but possibly wrong – cluster membership for some nodes.

   Navlakha, White, Nagarajan, Pop, and Kingsford, RECOMB 2009.

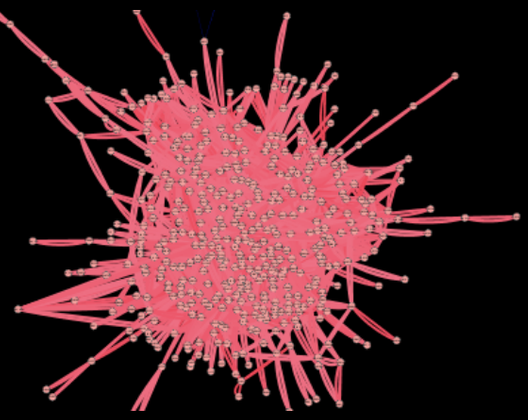
Clustering Probabilistic Graphs

Expected Quality Clustering

Clustering

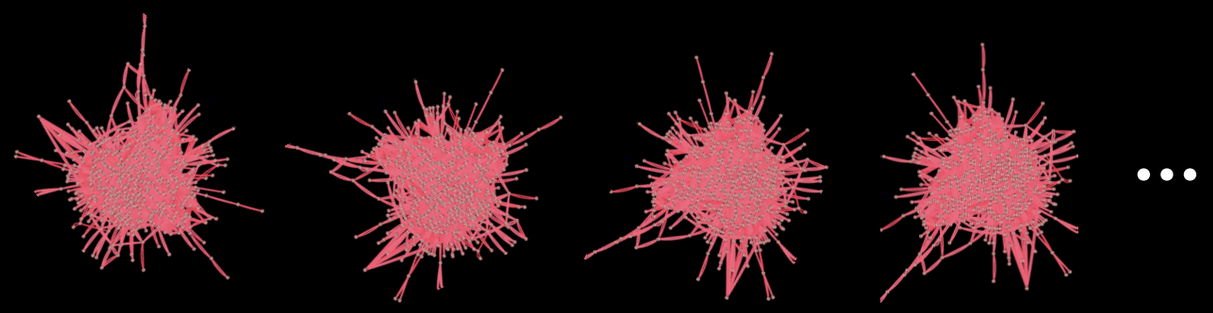
$$\operatorname{argmax}_{\mathcal{C}} f(G, \mathcal{C})$$

Clustering quality function



Single graph

$$\operatorname{argmax}_{\mathcal{C}} \mathbb{E}_G f(G, \mathcal{C})$$



Instantiations of the probabilistic graph.

Next: a heuristic for optimizing E f(G, C)

Acknowledgments

Current Group:

Guillaume Marçais*
Saket Navlakha
Justin Malin
Geet Duggal*
Darya Filippova

Darya not pictured

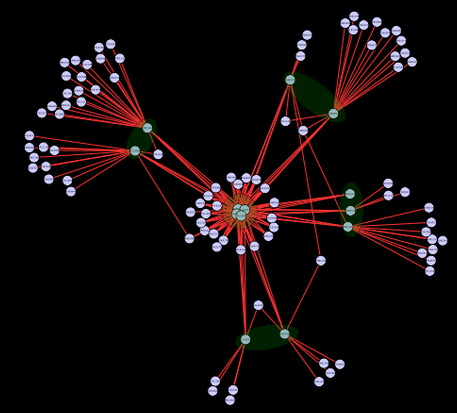
Former Students:

Grecia Lapizco-Encinas

C.K. was partially supported by NSF grants IIS-0812111 and EF-0849899.

Exploring Biological Network Dynamics with Ensembles of Graph Clusterings: *Pac. Symp. Biocomp.* (PSB) 2010.
Saket Navlakha, Carl Kingsford

Extracting between-pathway models from E-MAP interactions using expected graph compression: *RECOMB* 2010.
David E. Kelley, Carl Kingsford



CBCB
Center for Bioinformatics & Computational Biology
University of Maryland

Thanks!