# MITx: Statistics, Computation & Applications

Criminal Networks Module

Lecture 4: Applications Beyond Criminal Networks

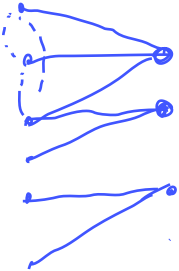
# Co-offender network

- All arrests in Quebec between 2003 and 2010
- Information on criminals and crime events they were arrested for
- Co-offender network: nodes are the offenders, and two offenders share a (possibly weighted) edge whenever they are arrested for the same crime event
- Summarizing the data of all arrests in Quebec as a network of co-offenders only portrays one side of the story

What information is lost in this representation of the data? What other representations are possible and what questions can be analyzed?

criminals

crimes



$A$

$AA^T$

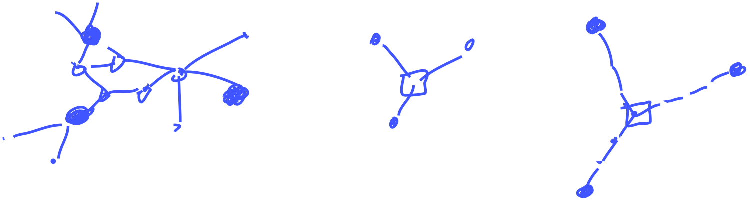
$A^TA$

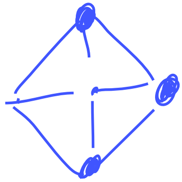
# Caviar network

- Investigation by Montréal police between 1994 and 1996
- Drug trafficking network investigated over time
- New criminals were added to the network by wire-tapping phones
- 11 seizures (money or drugs) throughout the investigation, but criminals were arrested only at the end
- Unique opportunity to analyze how a network reorganizes itself when subjected to stress

# Related scenario

- Given a social network and $k$ criminal suspects, how to determine other suspects?
- Same question is extremely important in biology: given certain genes that are known to cause a certain disease, determine other candidate genes (e.g. based on protein-protein interaction network for determining autism genes: http://dx.doi.org/10.1101/057828)





# MITx:
# Statistics, Computation & Applications

Criminal Networks Module

Lecture 4: Applications Beyond Criminal Networks

# Co-offender network

- All arrests in Quebec between 2003 and 2010
- Information on criminals and crime events they were arrested for
- Co-offender network: nodes are the offenders, and two offenders share a (possibly weighted) edge whenever they are arrested for the same crime event
- Summarizing the data of all arrests in Quebec as a network of co-offenders only portrays one side of the story

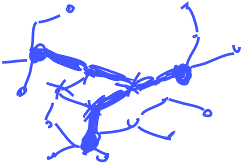
What information is lost in this representation of the data? What other representations are possible and what questions can be analyzed?

# Caviar network

- Investigation by Montréal police between 1994 and 1996
- Drug trafficking network investigated over time
- New criminals were added to the network by wire-tapping phones
- 11 seizures (money or drugs) throughout the investigation, but criminals were arrested only at the end
- Unique opportunity to analyze how a network reorganizes itself when subjected to stress

# Related scenario

- Given a social network and $k$ criminal suspects, how to determine other suspects?
- Same question is extremely important in biology: given certain genes that are known to cause a certain disease, determine other candidate genes (e.g. based on protein-protein interaction network for determining autism genes: http://dx.doi.org/10.1101/057828)
- How do we identify nodes that are "between" a given set of seed nodes?



# Steiner trees

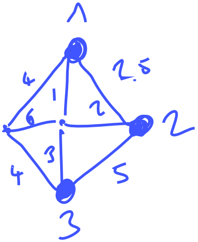
Determine a small subnetwork that contains the given suspects / genes and connects these nodes

# Steiner trees

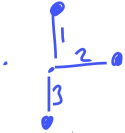
Determine a small subnetwork that contains the given suspects / genes and connects these nodes

**Steiner tree:**

- shortest subnetwork that contains a given set of nodes
- NP-complete problem
- polynomial time approximations:

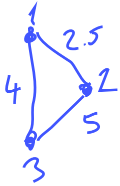


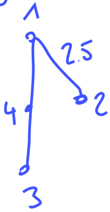

Steiner tree




weight 6

$$\begin{array}{c|c} & 1 \quad 2 \quad 3 \\ \hline \begin{array}{c}1\\2\\3\end{array} & \begin{bmatrix} & 2.5 & 4 \\ & & 5 \\ & & \end{bmatrix} \end{array}$$



# Steiner trees

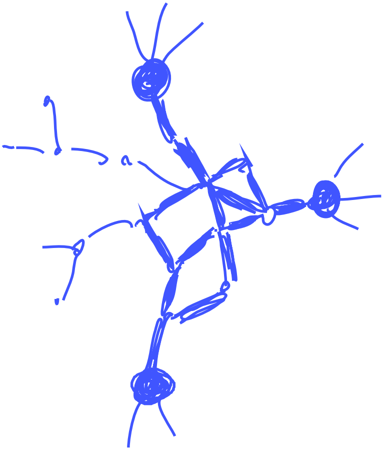
Determine a small subnetwork that contains the given suspects / genes and connects these nodes

**Steiner tree:**

- shortest subnetwork that contains a given set of nodes
- NP-complete problem
- polynomial time approximations:
  - compute minimum distance between the given set of nodes and determine minimum spanning tree in this new network $\rightarrow$ weight of resulting tree is within 2 times weight of optimal Steiner tree
  - best known approximation is of a factor of $\ln 4 + \epsilon < 1.39$ given by linear programming relaxation combined with iterative, randomized rounding (Byrka et al., 2013)

$\Rightarrow$ use collection of approximate Steiner trees for further analysis: autism interactome / criminal interactome

Genomics application: http://fraenkel-nsf.csbi.mit.edu/steinernet/tutorial.html



Interactome

# Analysis of autism interactome / criminal interactome

Is interactome indeed more tightly connected than at random?

- assume interactome was built with $k$ seed nodes
- choose $k$ nodes at random and compute resulting interactome
- perform hypothesis test based on diameter / average geodesic

$\Rightarrow$ compute nodes with high betweenness centrality in interactome to obtain candidate genes / suspects

# Co-offending network

- summarizing the data of all arrests in Quebec as a network of co-offenders only portrays one side of the story
- data can be represented by a binary matrix $A$ where the rows correspond to persons and the columns to crimes
- the co-offending network has (weighted) adjacency matrix $AA^T$
- similarly, we can build a network of crimes based on the (weighted) adjacency matrix $A^T A$
- or we could analyze the bipartite network given by the adjacency matrix

$$\begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}$$

**How do we go about detecting communities in these networks?**

# Community detection

## Community detection:

- detect subsets of nodes that are more densely connected between each other in the network than outside the community

## Clustering

- determine subsets of points that are 'close' to each other given a pairwise distance or similarity measure
- can be used also for community detection by defining a vertex similarity measure (e.g., geodesic distance, number of different neighbors, correlation between adjacency matrix columns, etc.)
- we already discussed some clustering methods in Module 1 (e.g. hierarchical clustering, k-means)

# Other methods: Divisive algorithm using betweenness

- Intuition: intercommunity edges have a large value of edge betweenness, because many shortest paths connecting vertices of different communities will pass through them
- Algorithm of Girvan and Newman (2002): iteratively remove edges with highest betweenness centrality
- can define betweenness using geodesic, flow or random walk

# Other methods: Modularity maximization

- quality function: function that assigns a number (quality measure) to each partition of a graph

# Other methods: Modularity maximization

- quality function: function that assigns a number (quality measure) to each partition of a graph
- most popular quality function: modularity

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - P_{ij})\, \delta(C_i, C_j),$$

where $P_{ij}$ is expected number of edges between $i$ and $j$ in null model

# Other methods: Modularity maximization

- quality function: function that assigns a number (quality measure) to each partition of a graph
- most popular quality function: modularity

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - P_{ij})\, \delta(C_i, C_j),$$

where $P_{ij}$ is expected number of edges between $i$ and $j$ in null model

  - compares actual edge density to expected edge density in null model
  - for Erdös-Renyi model $P_{ij} = \frac{2m}{n(n-1)}$
  - for configuration model $P_{ij} = \frac{k_i k_j}{2m-1}$
  - for bipartite graphs: $Q = \frac{1}{2m} \sum_i \sum_j (A_{ij} - \frac{k_i^{(1)} k_j^{(2)}}{2m})\, \delta(C_i, C_j)$,

# Louvain method (Blondel et al., 2008)

- modularity optimization is NP-complete (Brandes et al., 2006)
- Louvain method: very fast heuristic

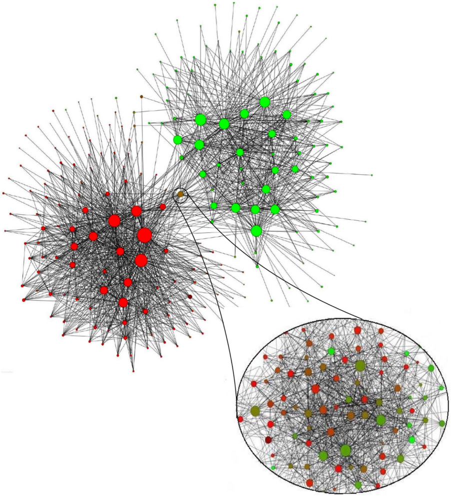
# Louvain method (Blondel et al., 2008)

- modularity optimization is NP-complete (Brandes et al., 2006)
- Louvain method: very fast heuristic
  - put each vertex in its own community
  - put vertex $i$ into community $j$ that yields biggest increase in modularity
  - replace communities by supervertices, where edge weight between supervertices is sum of edge weights between corresponding nodes
  - iterate process until $Q$ cannot be improved

# Louvain method (Blondel et al., 2008)

- modularity optimization is NP-complete (Brandes et al., 2006)
- Louvain method: very fast heuristic
  - put each vertex in its own community
  - put vertex $i$ into community $j$ that yields biggest increase in modularity
  - replace communities by supervertices, where edge weight between supervertices is sum of edge weights between corresponding nodes
  - iterate process until $Q$ cannot be improved
- provides decomposition of network into communities for different levels of organization
- extremely fast: runs in $\mathcal{O}(m)$
- can be applied to find communities in bipartite networks either using $AA^T$, $A^TA$ or $\begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}$

# Louvain method (Blondel et al., 2008)



Belgian mobile phone network with 2M customers (red: French-speaking, green: Dutch-speaking).

# References


- J. Byrka et al. *Steiner tree approximation via iterative randomized rounding*. Journal of the ACM 60, 2013
- S. S. Huang and E. Fraenkel. *Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks*. Science Signaling 2(81):ra40.
- G. Novarino, et al. *Exome sequencing links corticospinal motor neuron disease to common neurodegenerative disorders*. Science 343, 2014.
- A. Krishnan, et al. *Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder*. Nature Neuroscience, 2016.
- Lecture notes on the Laplacian and spectral clustering by Tim Roughgarden & Greg Valiant: `http://web.stanford.edu/class/cs168/l/l11.pdf`
- V. D. Blondel, et al. *Fast unfolding of communities in large networks*. Journal of Statistical Mechanics: Theory and Experiment 10, 2008.
- S. Fortunato. *Community detection in graphs*. Physics Reports 486, 2010.