Course project «Optimization approaches to community detection»

Marina Danilova, Alexander Podkopaev, Nikita Puchkin, Igor Silin

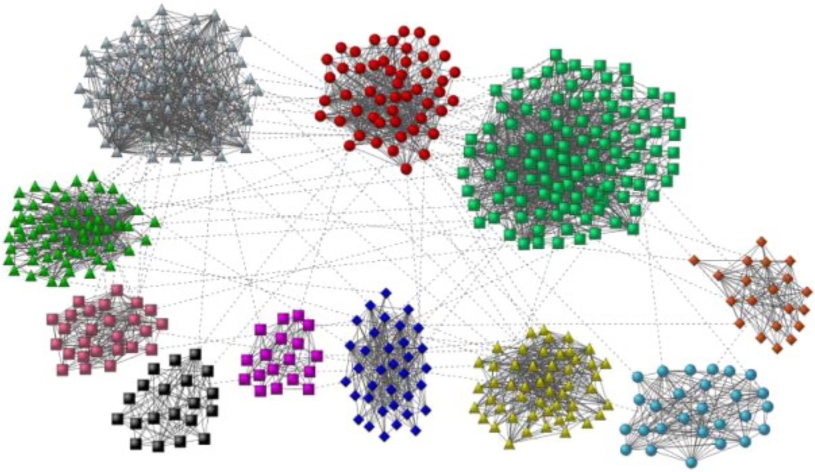
Skolkovo institute of science and technology

December 16, 2016

Plan

1. Introduction to community detection
2. Algorithms
    - Spectral method
    - Modularity-based method
    - Natural conjugate gradients method
    - Semidefinite relaxations
3. Experimental results

Example



Notations

Assumption

We consider **undirected unweighted** graphs **without loops** with n nodes. The nodes are enumerated as $\{1, \ldots, n\}$.
Graph is given by its $n \times n$ adjacency matrix A.

Goal of community detection

Find **partition** of nodes into **non-overlapping** clusters.
The number of clusters is k.
The clusters are denoted as $\{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$.

Subsets and Cuts

Measuring sizes of subsets

Let $\mathcal{C}_1, \dots, \mathcal{C}_k$ be subsets of vertices. Then:

- $|\mathcal{C}_i| = \{\text{number of vertices in } \mathcal{C}_i\}$
- $vol(\mathcal{C}_i) = \sum\limits_{i \in \mathcal{C}_i} d_i$

MinCut

Define $W(\mathcal{C}_p, \mathcal{C}_q) := \sum\limits_{i \in \mathcal{C}_p, j \in C_q} a_{ij}$. Then MinCut problem is:

$$cut(\mathcal{C}_1, \dots, \mathcal{C}_k) = \frac{1}{2} \sum_{i=1}^{k} W(\mathcal{C}_i, \overline{\mathcal{C}_i}) \to \min_{\mathcal{C}_1, \dots, \mathcal{C}_k}$$

Balancing Cuts

RatioCut and Normalized Cut

The MinCut solution separates one individual vertex from the rest. The following objectives are considered:

$$RatioCut(\mathcal{C}_1, \dots, \mathcal{C}_k) = \sum_{i=1}^{k} \frac{cut(\mathcal{C}_i, \overline{\mathcal{C}_i})}{|\mathcal{C}_i|} \to \min_{\mathcal{C}_1, \dots, \mathcal{C}_k}$$

$$Ncut(\mathcal{C}_1, \dots, \mathcal{C}_k) = \sum_{i=1}^{k} \frac{cut(\mathcal{C}_i, \overline{\mathcal{C}_i})}{vol(\mathcal{C}_i)} \to \min_{\mathcal{C}_1, \dots, \mathcal{C}_k}$$

Balancing conditions lead to NP-hard problem. Spectral clustering is a way to solve relaxed versions of those problems

Relaxed problem

Types of Laplacians

- Unnormilized Laplacian: $L = D - A$
- Symmetric Laplacian: $L_{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$
- Random walk Laplacian: $L_{rw} = D^{-1} L = I - D^{-1} A$

Idea

- Solving relaxed problem is equivalent to considering eigenvectors corresponding to k smallest eigenvalues of Laplacian that describe cluster properties of given graph

Related paper

Ulrike von Luxburg «A Tutorial on Spectral Clustering», 2007

Modularity-based method

Formulating an optimization problem

Modularity-based method

Natural conjugate gradients method

- Model parametrized by

  $z(i) \sim \text{Poly}(\pi), \quad i = \overline{1, n}$

  $P = \|p_{ij}\|_{i,j=\overline{1,k}}$ - probabilities of inter-cluster edges occurence

- Bayesian approach:

$$\pi \sim \text{Dirichlet}(\alpha)$$
$$p_{ii} \sim \text{Beta}(\beta), \quad i = \overline{1, k}$$
$$p_{ij} \ll 1, \quad \forall i \neq j$$

- $p(z, \pi, P|A)$ - true posterior with observed adjacency matrix A
- $\mathcal{Q}$ - family of feasible distributions

Natural conjugate gradients method

Formulating an optimization problem

$$\mathcal{L}(q) \equiv -\mathrm{KL}\left(q \| p(Z, \pi, P | A)\right) \longrightarrow \max_{q \in \mathcal{Q}}$$

- The problem can be reduced to unconstrained optimization:

$$\mathcal{L} = \mathcal{L}(\theta) \longrightarrow \max, \quad \theta \text{ - } n \times (k-1) \text{ matrix}$$

- Use conjugate gradients method in a statistical manifold
- Metrics is defined by a matrix

$$\mathcal{I}(\theta) - \text{Fischer information}$$

Semidefinite relaxations

Formulating an optimization problem

Semidefinite relaxations