Community Detection

fundamental limits & efficient algorithms

Laurent Massoulié, Inria
Community Detection

• From graph of node-to-node interactions, identify groups of similar nodes

Example: Graph of US political blogs’ citations [Adamic & Glance 2005]
Application 1: contact recommendation in online social networks

Data: “friendship” graph

→ recommend members of user’s implicit community

Variation. NSA’s “co-traveler programme”: Spot groups of suspect persons meeting regularly in unusual places
Application 2: item recommendation to users

Data: \{user-item\} matrix
Example: Netflix prize dataset $\rightarrow$ \{user-movie\} ratings

<table>
<thead>
<tr>
<th>User / Movie</th>
<th>Alice</th>
<th>?</th>
<th>**</th>
<th>***</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>***</td>
<td>?</td>
<td></td>
<td>?</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deirdre</td>
<td>*****</td>
<td>**</td>
<td></td>
<td>**</td>
</tr>
</tbody>
</table>

Item communities can guide recommendation: “users who liked this also liked...”
Application 3: categorizing chemical reactives in biology

Data: sets of chemicals and reactions involving them


More generally: **Knowledge graph** as generic representation of data
A1 has with B1 interaction of type C1
A2 has with B2 interaction of type C2
...

End goal: Algorithms with good accuracy at low computation cost

Outline:

– An algorithm

– Its performance when signal is strong

– Fundamental limits and better algorithms when signal is weak
Typical algorithm for community detection: first embed, then cluster
How to cluster

K-means clustering [Lloyd 1957]

Initialization: start with K centers placed at random
1) Cluster points according to their nearest center
2) Update center position to center of mass of associated points
How to cluster

K-means clustering [Lloyd 1957]

Initialization: start with K centers placed at random
1) Cluster points according to their nearest center
2) Update center position to center of mass of associated points
3) Iterate
How to cluster

K-means clustering [Lloyd 1957]

Initialization: start with K centers placed at random
1) Cluster points according to their nearest center
2) Update center position to center of mass of associated points
3) Iterate
Illustration in dimension 2 on Netflix dataset
How to embed:
The basic recipe for dimension reduction

- Karl Pearson’s Principal Components Analysis (PCA)
  “On Lines and Planes of Closest Fit to Systems of Points in Space”, 1901

Data vectors $z_1, \ldots, z_p$ in $n$-dimensional space

Linear Algebra ahead! Data matrix: $Z = [z_1 | z_2 | \ldots | z_p]$

$D$-dimensional subspace that best approximates data vectors:

Obtained from eigenvectors $x_1, \ldots, x_D$ of $ZZ^T$ corresponding to its $D$ largest eigenvalues
Spectral Embedding

- Data representation by adjacency matrix $A$ of graph:

![Adjacency matrix example](image)

$A_{uv}^t$ encodes paths in graph: $A_{uv}^t$ = number of paths of length $t$ from $u$ to $v$

$A^2 = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 2 & 1 & 1 \\ 1 & 1 & 2 & 0 \\ 1 & 1 & 0 & 2 \end{pmatrix}$

- $(\text{eigenvector}, \text{eigenvalue}) (x, \lambda)$ pair of $A$ verifies for all $t$:

$$\lambda^t x_v = \sum_u x_u \times \text{number of paths of length } t \text{ from } u \text{ to } v$$
Spectral Embedding

“Principal Components Analysis”:

From matrix $A$, extract $D$ normed eigenvectors $x_1, \ldots, x_D$ corresponding to $D$ largest eigenvalues $|\lambda_1| \geq \cdots \geq |\lambda_D|$

$\rightarrow$ Vectors $\hat{z}_u$ in $D$-dimensional space closest to column vectors of $A$:
$$\hat{z}_u = x_1(u)\lambda_1 x_1 + \cdots + x_D(u)\lambda_D x_D$$

$\rightarrow$ Spectral embedding: form $D$-dimensional node representatives
$$y_u = \{x_i(u)\}_{i=1\ldots D}$$
How good is this method?
Should we replace adjacency by other matrix?
How do amount & quality of data affect achievable accuracy?
The need for generative models of data

Empirical comparison of algorithms on specific datasets: necessary, but
– provides only limited understanding of their merits

The problem of ground truth:
Where are the true Democrats?

Analysis of algorithms on data from generative model:
– enables to quantify quality of algorithms
– reveals fundamental limits on feasibility of community detection
– guides design of new algorithms
The Stochastic Block Model
[Holland-Laskey-Leinhardt’83]

- Nodes in block 1

![Diagram of nodes in blocks 1, 2, and 3 with edges between them.]

- \( n \) nodes, partitioned into blocks
- Edge between nodes \( u, v \) present at random with probability depending only on their blocks \( k(u), k(v) \)
- \( n \gg 1 \)
Schematic view of community detection

Signal: node blocks $\kappa(u)$
Alternatively, block matrix of edge probabilities

Observation: adjacency matrix $A$

Extraction of signal: estimators $\hat{\kappa(u)}$ of node blocks,
Accuracy: fraction of correctly classified nodes

→ A high-dimensional statistical inference problem
Efficiency of spectral approach in a strong-signal regime

For edge probabilities $P(u \sim v) = \frac{d}{n} \times F_{k(u)k(v)}$ with fixed parameters $F_{ij}$, Factor $d$: measures signal strength; strong signal: $d \gg 1$
Efficiency of spectral approach in a strong-signal regime

Non-zero eigenvalues of signal matrix appear nearly unchanged
Signal captured in their eigenvectors

Spectrum of adjacency matrix, strong-signal case
One word on random matrices

Study initiated by Eugene Wigner (1955)

Wigner’s semi-circle law [Wigner’55]:

Spectrum of symmetric $n \times n$ matrix with random Gaussian entries with zero mean and variance $\frac{\sigma^2}{n}$ is supported in $[-2\sigma, 2\sigma]$, with asymptotic distribution
Efficiency of spectral approach in a strong-signal regime

• Noise matrix in our observations: elements of variance $O\left(\frac{d}{n}\right)$:

  → eigenvalues of order $O(\sqrt{d})$ when $d = \Omega(\ln n)$; [Feige-Ofek 2005]
  (a result expected in view of Wigner’s semi-circle law)

Spectrum of signal matrix $\begin{bmatrix} \square & \square & \square \\ \square & \square & \square \\ \square & \square & \square \end{bmatrix}$ : eigenvalues of order $d$

→ For $d = \Omega(\ln n)$, spectrum of noise negligible compared to spectrum of signal
→ Spectral method correctly clusters all but a vanishing fraction of nodes
  (by results on perturbation of eigenvalues and eigenvectors)
Efficiency of spectral approach in a strong-signal regime

Non-zero eigenvalues of signal matrix appear nearly unchanged
Signal captured in their eigenvectors

Spectrum of adjacency matrix, strong-signal case

Weaker signal: useful information, if any remains, is no longer concentrated in a few eigenvectors
Fundamental limits to community detection: 
Low signal regime, $d = O(1)$

- The insight from statistical physics: 
  [Decelle-Krzakala-Moore-Zdeborova 2011] Conjecture

- There is a phase where the observations contain no information, and no estimators $\hat{k}(u)$ can do better than random guess:
  
  Community Detection is information-theoretically impossible

- There is a phase where better-than-random detection can be achieved in polynomial-time

  Community Detection is feasible from both informational and computational viewpoints
Fundamental limits to community detection: Low signal regime, $d = O(1)$

- Illustration in a symmetric two-communities scenario:

- For $\tau = \frac{(a-b)^2}{2(a+b)} \leq 1$, no estimator $\hat{k}$ can do better than random guess (1/2 of nodes misclassified)

→ Below this threshold, CD is information-theoretically impossible

- For $\tau > 1$, better-than-random detection can be achieved in polynomial-time

→ Above this threshold, CD is feasible from both informational and computational viewpoints
The argument for feasibility: fixing the spectral method

- First approach (LM’13): consider instead matrix $S$ where $S_{uv}$: number of self-avoiding walks of length $t$ in graph connecting $u$ to $v$

→ “Nice” spectrum for suitable $t$: eigenvectors enable better-than-random node classification whenever $\tau: = \frac{(a-b)^2}{2(a+b)} > 1$

→ Polynomial-time, but counting self-avoiding paths is cumbersome
Alternative: “Spectral Redemption”  

- Non-backtracking matrix $B$:
  Defined on oriented edges $\overrightarrow{uv}$ for $(u, v) \in E : B_{uv,xy} = 1_{v=x} 1_{u \neq y}$

→ Asymmetric, such that $B_{ef}^k = \text{number of non-backtracking paths on } G \text{ of length } k+1 \text{ starting at } e \text{ and ending at } f$

Method: obtain leading eigenvectors of $B$ and project them into node-indexed vectors to perform embedding
Spectrum of non-backtracking matrix, stochastic block model

Circle of radius $\sqrt{\mu_1}$

$\lambda_2 \approx \mu_2$

$\lambda_1 \approx \mu_1$

$\mu_i$ : eigenvalues of signal matrix
Non-backtracking spectra of stochastic block models [Bordenave-Lelarge-LM, 2015]

- $\mu_1, \ldots, \mu_r$, $|\mu_1| \geq \cdots \geq |\mu_r|$ : eigenvalues of signal matrix

→ If $\mu_i^2 > \mu_1$ then $B$ has eigenvalue $\lambda$ close to $\mu_i$ and corresponding eigenvector is correlated with underlying blocks

The rest of $B$'s spectrum lies in the disk $\{|z|^2 \leq \mu_1\}$

Implies better-than-random detection feasible in polynomial time whenever there exists $i > 1$ such that $\mu_i^2 > \mu_1$ as predicted in [KMMNSZZ’13]

→ The so-called Kesten-Stigum condition, which generalizes condition $\tau > 1$ to more than two communities
Spectrum of non-backtracking matrix, political blogs
Conjectured phase diagram for community detection at low signal

(assuming fixed inter-community parameter $b$)

- Number of blocks $K$
- Intra-block parameter $a$

Above Kesten-Stigum threshold:
- Detection “easy”
  (spectral methods++)

Detection feasible but hard:
- (no known polynomial time algorithm)

Detection infeasible

$K = 2$
Conclusions

- Community Detection motivates search for new algorithms
  
  → Led to spectral methods with self-avoiding & non-backtracking path counts, but others are yet to be invented

- Community Detection in Stochastic Block Model: rich playground for analysis of computational complexity with methods of statistical physics and probability theory
  
  → What can be said about the hard phase???
BACKUP
Spectrum of non-backtracking matrix
Erdős-Rényi graph (1 community)
The argument for impossibility
[Mossel-Neeman-Sly 2012]

• An easier problem: predicting the block $k(u)$ of some node, if one were given the blocks $k(v)$ of nodes $v$ at some large graph distance

→ This corresponds to so-called tree reconstruction problem: predict trait of ancestor from observed traits of far-away descendants

→ Phase transition on feasibility of tree reconstruction characterized in [Evans-Kenyon-Peres-Schulman’00]
Ramanujan graphs
[Lubotzky-Phillips-Sarnak’88]
Corollary:
Erdős-Rényi graphs are nearly Ramanujan
Open questions for detection in SBM’s (2)