EDGE LABEL INERENCE IN GENERALIZED STOCHASTIC BLOCK MODELS

Jiaming Xu (UIUC)
Laurent Massoulié (MSR-Inria Joint Centre, Inria)
Marc Lelarge (Inria)
The Stochastic Block Model
[Holland-Laskey-Leinhardt’83]

- $n$ “nodes” partitioned into $K$ categories
- Category $k$: $\alpha_k \ n$ nodes
- Edge between nodes $u,v$ present with probability $b_{k(u)k(v)} \ s/n$
  
$s$: “signal strength”

$\rightarrow$ Observation: adjacency matrix $A$
SBM with labels

- Edges \((u-v)\) labeled by \(L_{uv} \in L\) (finite set)
- Drawn from distribution \(\mu_{k(u)k(v)}\)

E.g. Netflix ratings: labels 1-5 stars
SBM with general types
[Bollobas-Janson-Riordan’07,…]

- User type $k(u)$ i.i.d. in general set (e.g. uniform on $[0,1]$)
- Edge $(u-v)$ present w.p. $b_{k(u)k(v)} s/n$ for “kernel” $b$
  e.g. $b_{x,y} = F(x - y)$

- Edges $(u-v)$ labeled by $L_{uv} \in L$ (finite set)
- Drawn from distribution $\mu_{k(u)k(v)}$
Pre-processing: random label projections

Form matrix $\{A_{ij}W(L_{ij})\}$ from random (iid uniform [0,1]) projections $W(l)$ of labels
Spectral properties for logarithmic $S$

[tools: Kolchinski'98, Feige-Ofek’05]

- Define integral operator $Tf(x) := \int [\sum_l W(l)\mu_{xy}(l)f(y)P(dy)]$
- spectrum of $s^{-1}\{A_{ij}W(L_{ij})\} \approx$ spectrum of $T$

- Flexible model
- power-law spectra
- better matches to Netflix data
Label inference for logarithmic $s$

- Spectral embedding:

  form $R$-dimensional node representatives

  $$\mathbf{y}_u = \sqrt{n} \left\{ \frac{\lambda_k}{\lambda_1} x_k(u) \right\}_{k=1 \ldots R}$$

  \(\rightarrow\) Captures geometry of hidden node types $k(u)$ with accuracy controlled by “residual energy”

  $$\sum_{k>R} \lambda_k^2$$

  of operator’s spectrum
Consistent estimation of label distributions
Consistent estimation of label distributions

$\text{Prob}(\text{label}(i,j)=5)$

Node $j$

Node $i$
Consistent estimation of label distributions

\[ \text{Prob}(\text{label}(i,j)=5) \]

Use empirical distribution of labels \( L(i,k) \) for \( k \) in neighborhood of \( j \)
For more (including impossibility results at $O(1)$ signal strength): check poster & paper