Consistency of co-clustering for exchangeable array data

David Choi
(joint work with Patrick Wolfe, UCL)

Carnegie Mellon University

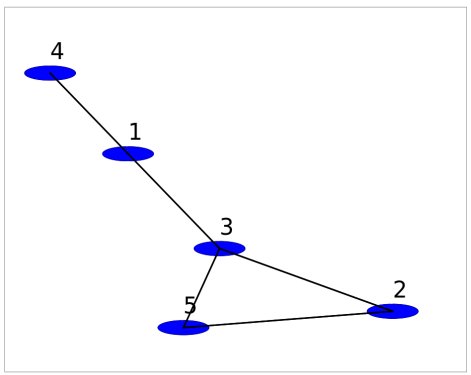
Oct 6, 2013

What is a network?

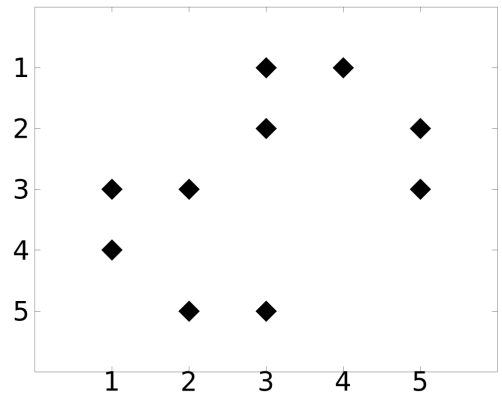
Network data records **interactions** between units, such as

- Communication/friendships/etc. between people
- Chemical reactions between proteins

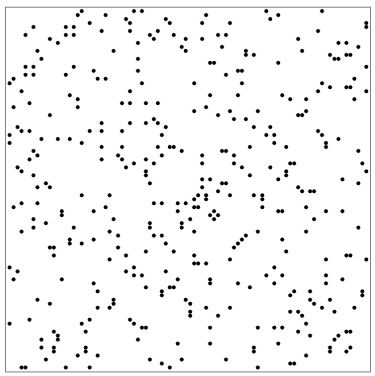
Common depictions for network data:



Graph



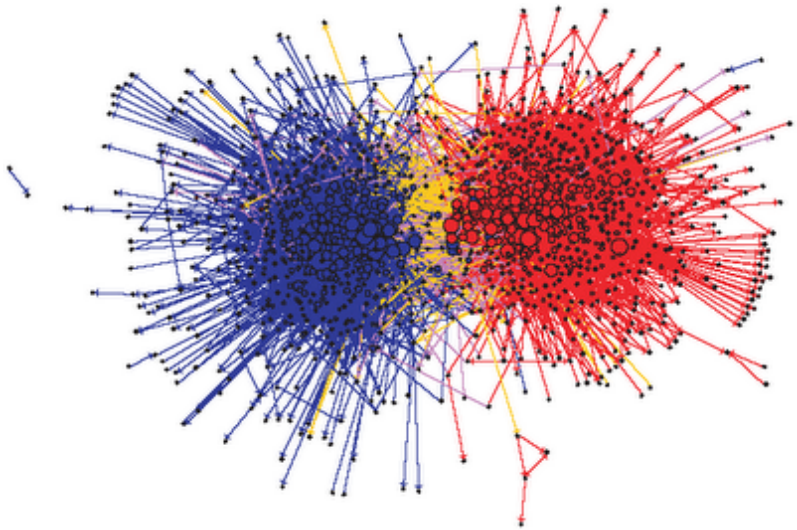
Adjacency matrix



Large adj. matrix

Identifying communities in networks

Clustering + post-hoc checks to find community structure

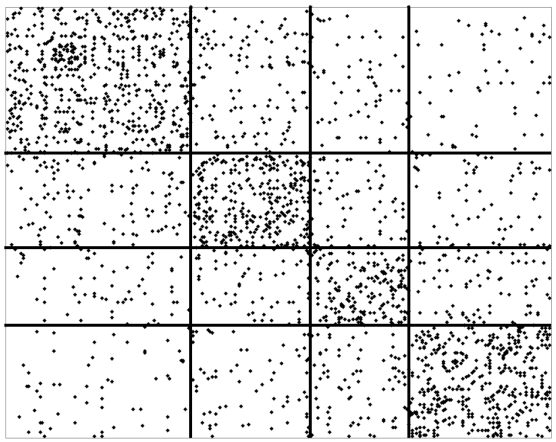


Political blog network, Adamic 2005

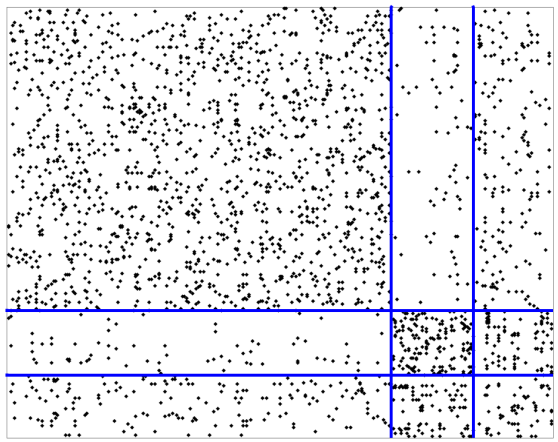
- Network equivalent of **clustering**
- **Post-hoc checks** are usually required to interpret clusters
- Checks may be effort-intensive – should be reserved for **statistically significant** findings
- But statistical significance is **not well understood** for community detection

Example from Ad-Health dataset

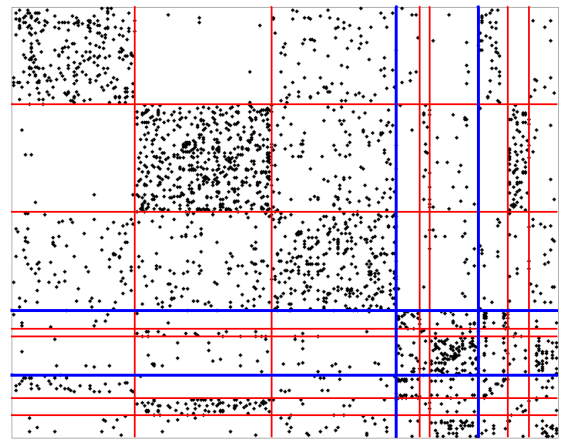
Survey of high school friendships



students grouped by **school year** (black lines)



students grouped by **race** (blue lines)



students grouped by **race** (blue lines) and within race by **likelihood-based clustering** (red lines)

Could this be due to noise?

What is known about community detection

Assuming a particular parametric generative model (stochastic blockmodel):

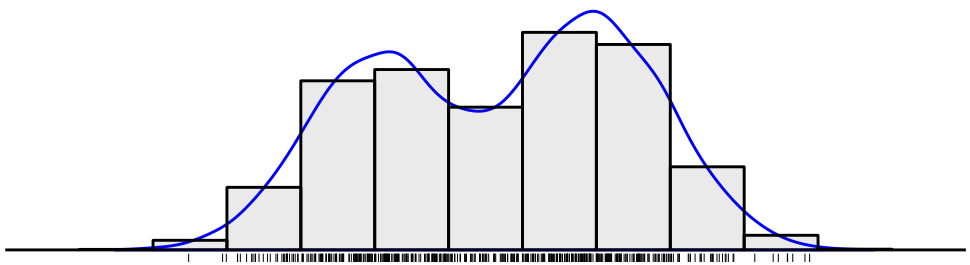
- Polynomial time algorithm (spectral clustering) finds correct clusters, with fraction of errors $\to 0$
- Exhaustive search (profile likelihood) finds correct clusters, with total number of errors $\to 0$
  - Also yields asymptotically normal parameter estiamtes

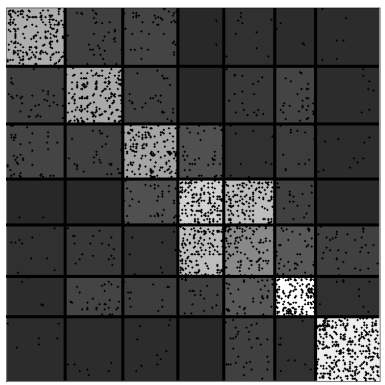
Issue: Assuming a stochastic blockmodel seems unrealistic – what if "true communities" don't exist in reality?

Main result

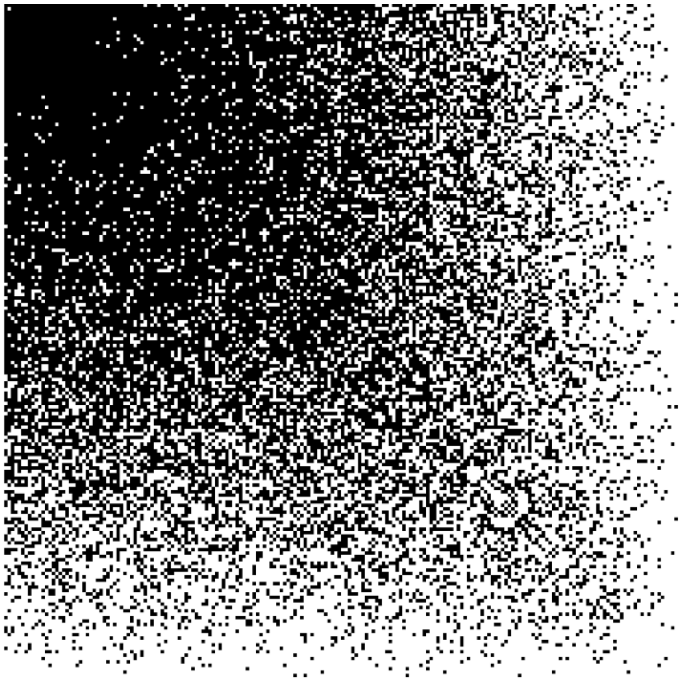
- Assume network data is generated a general nonparametric model, and then fit by a community detection model known as "stochastic co-blockmodel"
- Then the resulting estimate is a near-optimal "piecewise constant" approximation to the generative model

Implication: detected communities are consistent, in a sense analogous to **histograms**





Idea of nonparametric model



Adjacency matrix

Model $\omega_\alpha : [0,1]^2 \mapsto [0,1]$

Figures from Lovasz, L. "Very Large Graphs" (2013)

Description of nonparametric model

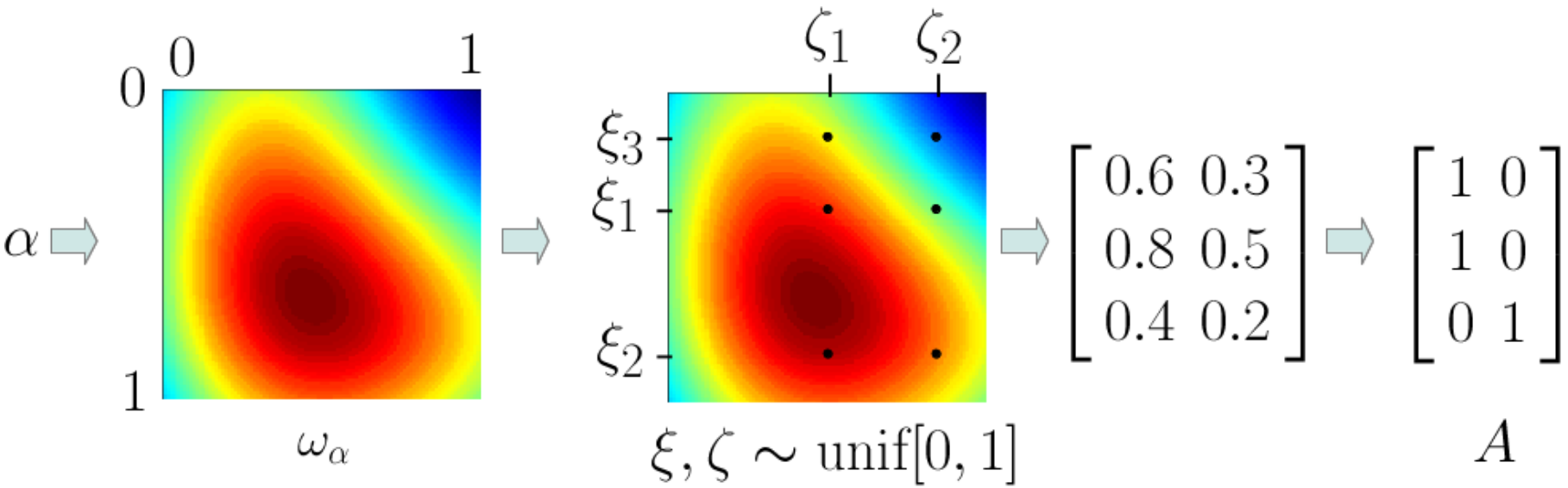
- ω_α maps $[0,1]^2 \mapsto [0,1]$, parameterized by α
- The directed adjacency matrix A is generated as follows:

$$\begin{aligned}
\alpha &\sim \text{Unif}[0,1], \\
\xi_i &\overset{\text{iid}}{\sim} \text{Unif}[0,1] \\
\zeta_i &\overset{\text{iid}}{\sim} \text{Unif}[0,1] \\
A_{ij}|\xi_i,\zeta_j &\overset{\text{iid}}{\sim} \text{Bernoulli}(\omega_\alpha(\xi_i,\zeta_j))
\end{aligned}$$

- ω_α cannot be distinguished from $(x,y) \mapsto \omega_\alpha(\pi_1(x),\pi_2(y))$ for any measure preserving transformations π_1, π_2.

Interpretation: ξ_i, ζ_i are latent factors controlling i's propensities in sending and receiving links

Picture of generative process



- ξ and ζ are latent
  - Otherwise estimation of ω_α would be straightforward
- $\text{Eq}(\omega_\alpha)$: equivalence class of ω_α, induced by measure-preserving maps on $[0, 1]$
- Mapping $\alpha \mapsto \omega_\alpha$ is not identifiable from a single network

Note: results will not require ω_α to be smooth, only measurable

A natural model for exchangeable data

Model encompasses all infinitely exchangeable arrays

Definition (Exchangeable arrays)

Binary array $\{A_{ij}\}_{i,j=1}^{\infty}$ is *separately exchangeable* if

$$\mathbb{P}(A_{ij} = X_{ij},\ i,j \in [n]) = \mathbb{P}(A_{ij} = X_{\eta_1(i)\eta_2(j)},\ i,j \in [n])$$

for all n, all permutations η_1, η_2 of $[n]$, and all X.

Theorem (Aldous-Hoover)(di Fenetti)

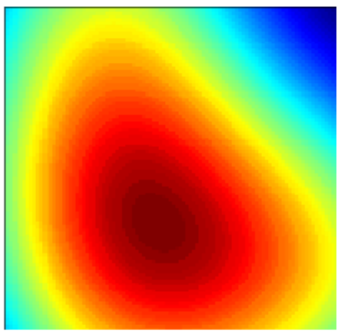
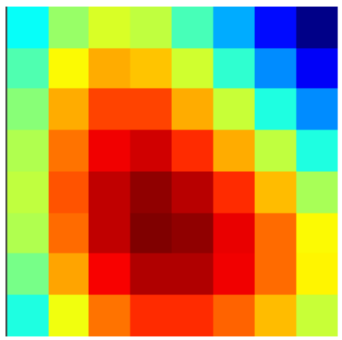
Let $\{A_{ij}\}_{i,j=1}^{\infty}$ be a separately exchangeable binary array. There exists $\underset{\sim}{\omega}$ which generates A.

Given a static snapshot of a network, natural to require the model to be invariant to a permutation of the adjacency matrix

We'll discuss later: are sparse graphs exchangeable?

Approximating ω_α by a piecewise constant function

Question: Can we fit a piecewise constant approximation to the generative model?

Given $\begin{bmatrix} 1 & 0 & \dots \\ 1 & 0 & \\ \vdots & & \end{bmatrix}$ generated by  , estimate 

A ω_α ω_ϕ

Parameters $\phi \equiv (\mu, \nu, \theta)$ describe ω_ϕ:

- Vectors μ, ν : boundaries of the piecewise constant regions (i.e., nonuniform grid)
- Matrix θ: heights of the piecewise constant regions

Fitting criteria for approximation

Fit by finding optimal clustering

We can fit $\phi = (\mu, \nu, \theta)$ to an observed adjacency matrix A by various criteria:

$$\text{Likelihood: } L_A(\mu, \nu, \theta) = \max_{S,T} \sum_{i,j} \log \mathbb{P}(A_{ij} \mid \theta_{S(i)T(j)})$$

$$\ell_2 \text{ error: } R_A(\mu, \nu, \theta) = \min_{S,T} \sum_{i,j} \left(A_{ij} - \theta_{S(i)T(j)}\right)^2$$

Mappings S and T assign nodes to K clusters, and are constrained to have assignment proportions matching μ and ν.

Metrics for approximation quality

In what sense might ω_ϕ be close to ω_α?

Fitting criteria correspond to risk functions that measure distance to the unknown ω:

$$L_\omega(\phi) = \inf_{\omega \in \mathrm{Eq}(\omega_\alpha)} \mathbb{E}\,\mathrm{KL}(\omega(\xi,\zeta), \omega_\phi(\xi,\zeta))$$

$$R_\omega(\phi) = \inf_{\omega \in \mathrm{Eq}(\omega_\alpha)} \mathbb{E}(\omega(\xi,\zeta) - \omega_\phi(\xi,\zeta))^2$$

$\mathrm{Eq}(\omega_\alpha)$: equivalence class of ω_α (all measure preserving maps of $[0,1]$)

KL: Kullback-Leibler divergence

First main result

Excess risk goes to zero

Theorem

For $\bar{\phi}$ maximizing[1] the true L_ω, and $\hat{\phi}$ maximizing its proxy L_A, such that $\bar{\theta}$ and $\hat{\theta}$ are bounded away from 1 or 0,

$$L_\omega(\bar{\phi}) - L_\omega(\hat{\phi}) = O_P\left(\frac{1}{n^{1/4}}\right).$$

Also, for $\bar{\phi}$ minimizing R_ω, and $\hat{\phi}$ minimizing R_A,

$$R_\omega(\hat{\phi}) - R_\omega(\bar{\phi}) = O_P\left(\frac{1}{n^{1/4}}\right).$$

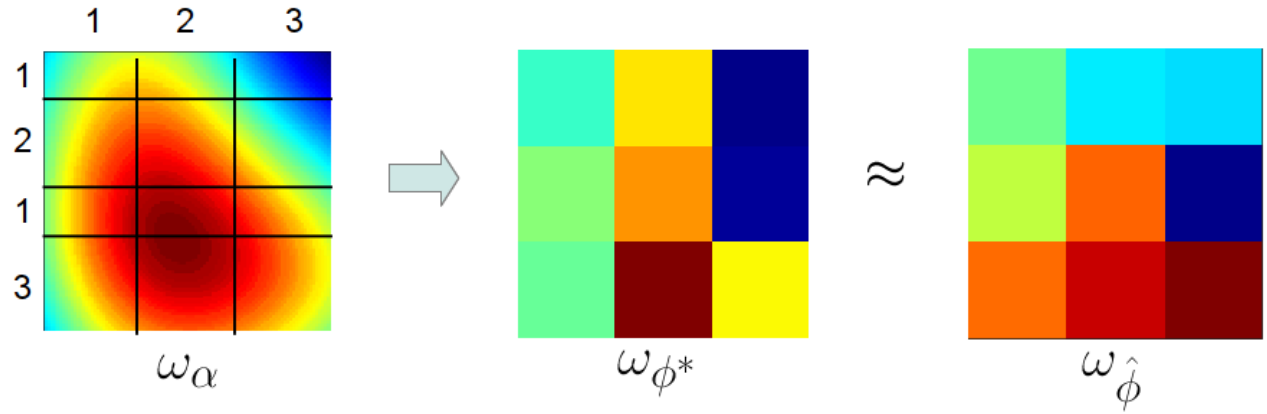
The estimate will be asymptotically optimal in terms of minimizing risk.

[1] over class frequencies μ, ν which are multiples of $1/n$

Significance of detected community structure

Corollary of Theorem 1

The estimate $\hat{\phi}$ approximates ϕ^*, for some some blockmodel ϕ^* induced by partitioning ω_α:



By approximate we mean:

- Identical region boundaries μ and ν
- Heights of piecewise constant regions are close: $\|\theta^* - \hat{\theta}\|_2^2 = O_P\left(\frac{1}{n^{1/4}}\right)$.

Implication: if you see an extreme partition in the data, then one also exists in the model

Discussion (1/2)

Why new approach was needed

- In standard learning problems, we want to choose ϕ to minimize expected loss compared to ω_α:

$$R_\omega(\phi) = \mathbb{E}(\omega_\alpha(\xi, \zeta) - \omega_\phi(\xi, \zeta))^2$$

- We can't evaluate R_ω, so we use noisy samples instead:

$$R_A(\phi) = \frac{1}{n^2} \sum_{i,j} (A_{ij} - \omega_\phi(\xi_i, \zeta_j))^2$$

- We'd try to show that $R_A(\phi) \approx R_\omega(\phi)$ uniformly over ϕ
  - Note: the n^2 samples (ξ_i, ζ_j) are not independent. Better to think of as n independent samples.

Discussion (2/2)

Why new approach was needed

What happens now that (ξ_i, ζ_j) are latent?

- ω_α is indistinguishable under measure-preserving maps:

$$R_\omega(\phi) = \inf_{\omega \in \mathsf{Eq}(\omega_\alpha)} \mathbb{E}(\omega(\xi, \zeta) - \omega_\phi(\xi, \zeta))^2$$

- We have to estimate ϕ and the latent ξ, ζ jointly:

$$R_A(\phi) = \min_{\xi_i, \zeta_j} \frac{1}{n^2} \sum_{i,j} (A_{ij} - \omega_\phi(\xi_i, \zeta_j))^2,$$

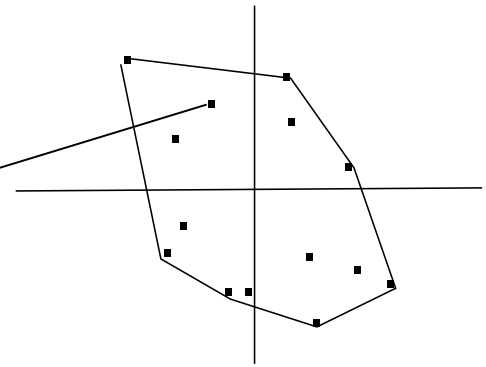
  Only n samples – won't suffice for union bound

Our approach: optimization over ξ_i, ζ_j has additional structure

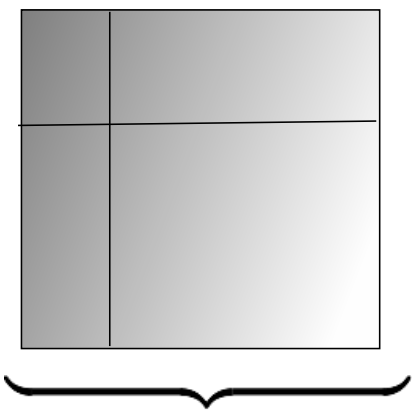
Second main result

Convex hulls of partition spaces $\mathcal{F}_A$ and $\mathcal{F}_\omega$

$$\underbrace{\left[\begin{array}{cc|c|c} 1 & 0 & 0 & \\ 1 & 1 & 0 & \cdots \\ \hline 0 & 0 & 1 & \\ & \vdots & & \ddots \end{array}\right]}_{A} \qquad \underbrace{\begin{bmatrix} 3 & 0 \\ 13 & 24 \end{bmatrix}}_{A/ST}$$

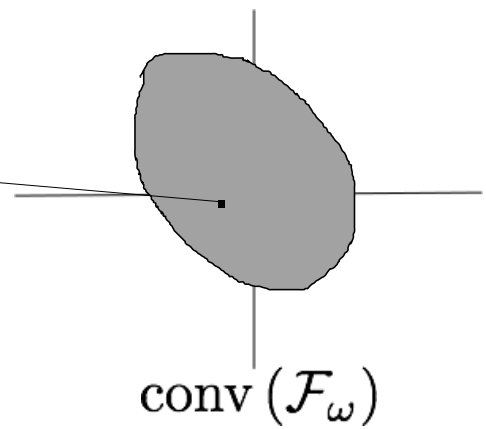


$$\operatorname{conv}(\mathcal{F}_A)$$



$$\underbrace{\qquad\qquad}_{\omega}$$

$$\underbrace{\begin{bmatrix} .03 & .06 \\ .02 & .1 \end{bmatrix}}_{\omega/\sigma\tau}$$

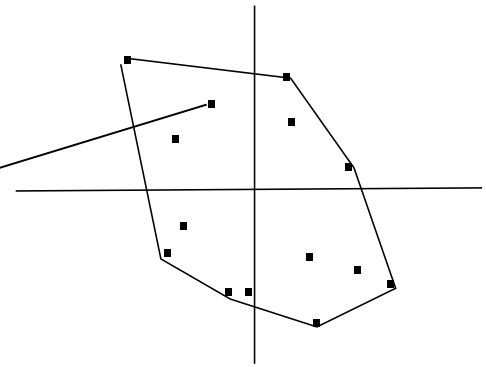


$$\operatorname{conv}(\mathcal{F}_\omega)$$

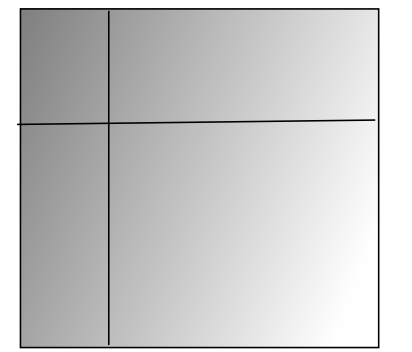
Second main result

Convex hulls of partition spaces $\mathcal{F}_A$ and $\mathcal{F}_\omega$

$$\underbrace{\left[\begin{array}{cc|c|c} 1 & 0 & 0 & \\ 1 & 1 & 0 & \cdots \\ \hline 0 & 0 & 1 & \\ & \vdots & & \ddots \end{array}\right]}_{A} \qquad \underbrace{\begin{bmatrix} 3 & 0 \\ 13 & 24 \end{bmatrix}}_{A/ST}$$

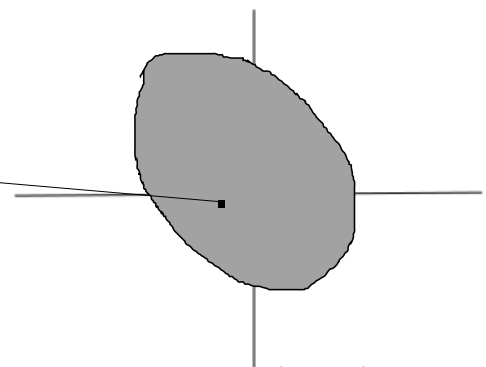


$\operatorname{conv}(\mathcal{F}_A)$



ω

$$\underbrace{\begin{bmatrix} .03 & .06 \\ .02 & .1 \end{bmatrix}}_{\omega/\sigma\tau}$$

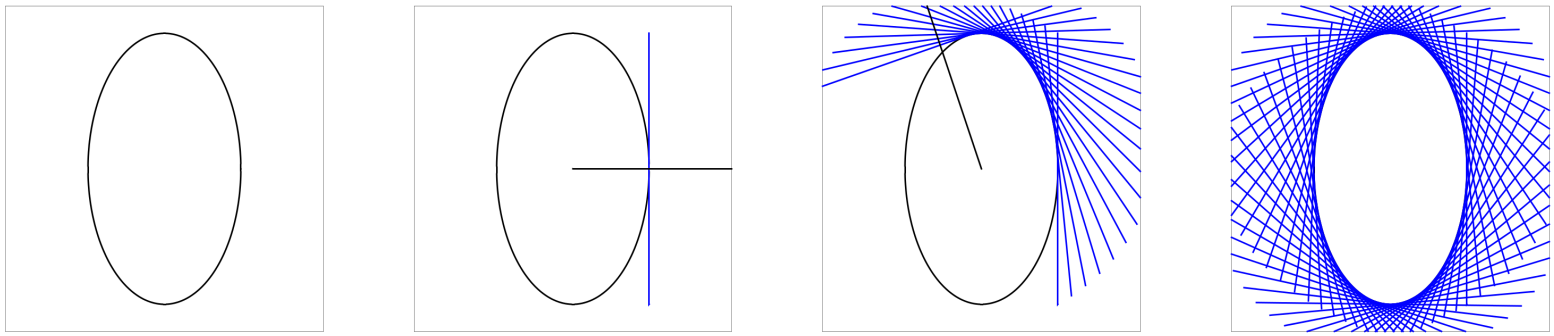


$\operatorname{conv}(\mathcal{F}_\omega)$

Theorem: $\max_{\mu,\nu} \; d_{\text{Hausdorff}}\left(\operatorname{conv}(\mathcal{F}_\omega), \operatorname{conv}(\mathcal{F}_A)\right) = O_P\left(\frac{1}{n^{1/4}}\right)$.

Relation to excess risk bound (1/2)

Prelude: understanding convex hulls



- Notion of **supporting hyperplane** in each direction θ
- The set of supporting hyperplanes in all directions induces the convex hull

Relation to excess risk bound (2/2)

Simple algebra (expanding the square):

$$R_A(\mu,\nu,\theta) = \min_{S,T} \sum_{i,j} \left(\theta_{S(i)T(j)} - A_{ij}\right)^2,$$

$$= \sum_{i,j} \theta^2_{S(i)T(j)} - 2 \underbrace{\max_{S,T} \langle \theta, A/ST \rangle}_{\text{supporting hyperplane}} + \sum_{i,j} A^2_{ij}$$

- $R_A(\theta)$ is determined by boundary of convex hull of $\mathcal{F}_A$ in direction θ
- Similar result relates R_ω to $\text{conv}(\mathcal{F}_\omega)$.
- Convergence of convex hulls $\rightarrow$ convergence of risk functions

Intuition of proof

- The supporting hyperplane

$$\max_{S,T} \langle \theta, A/ST \rangle$$

  involves an optimization over assignments S, T of all N nodes to clusters
- Show if $\sqrt{N}$ nodes are assigned optimally, and remainder are assigned "greedily", clustering is near-optimal[2]
- This implies the intrinsic dimensionality is $\sqrt{N}$
- Hence, N samples will suffice to estimate supporting hyperplane

[2] Alon et. al. "Random sampling and approximation of MAX-CSPs." 2003.

Relation to graph limits literature

- Approach is taken from Borgs et. al. (2008) work on graph limits
  - Defines convergence of dense graphs to be convergence of all subgraph frequencies
  - Closure (i.e. the limit objects) shown to be ω_α, termed "graphons"
  - Proved law of large numbers under this sense of convergence
- Relation to this work (roughly):
  - They show $\mathcal{F}_A \to \mathcal{F}_\omega$, instead of $\text{conv}(\mathcal{F}_A) \to \text{conv}(\mathcal{F}_\omega)$.
  - Their proof requires Szemeredi lemma and has exponentially slower rate

- Multiple notions of convergence exist for sparse graphs, with closure not yet known

Dense vs. sparse graphs

Is exchangeablity assumption always appropriate?

- Exchangeability is bad for modeling asymptotics when graph density $\to 0$.
  - Assumes that ω_α is fixed with n
  - One approach: assume model is $\rho_n \omega_\alpha$, where $\rho_n \to 0$ and normalize risk by ρ_n
    - Our results carry over; risk bound of $O_P(1/\sqrt{d})$ if degree $d = \omega(\log n)$
    - Issue: model becomes very smooth for large n

- For finite data, when is the sparse asymptotic regime appropriate?
  - Opinion: issue is one of enforcing **sparsity** and **connectedness** at once
    - If average degree is d, then expected number of isolates under ω_α is at least e^{-d}

Conclusions

Summary

- Blockmodels may be useful for exploratory data analysis of network data
- Even if generative model is not even approximately a blockmodel
- Akin to taking histograms

Future work

- Symmetric graphs
- Sparse graphs with $\omega(\log n)$ degree.
- Nonparametric estimation: letting model complexity increase with n so that total risk $\to 0$.
- Conjecture: proof techniques may be useful for other latent variable models.

Backup slides

Description of co-blockmodel

Let μ, ν be points on K-simplex. Let $\theta \in [0,1]^{K \times K}$.

Stochastic co-blockmodel

Given μ, ν, θ, the graph $G = (V_1, V_2, E)$ and its adjacency matrix $A \in \{0,1\}^{m \times n}$ are generated as follows:

1. *Generate latent variables $S = (S_1, \ldots, S_m) \sim \mu$, and $T = (T_1, \ldots, T_m) \sim \nu$*
2. *Let A_{ij} be Bernoulli with parameter $\theta_{S(i)T(j)}$. Connect (i, j) if $A_{ij} = 1$.*

Intuition: Discrete number of classes, with class probabilities μ and ν, and connections probabilities θ.

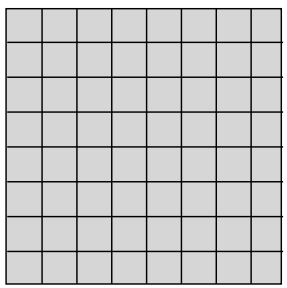
Note: Only identifiable up to label-switching.

Picture of near-optimal clustering

Similar to Borgs et al. (2008), Alon et al. (2003)

$$\max_{S,T} \langle \theta, A/ST \rangle = \max_{ST} \sum_{i=1}^{m} \sum_{j=1}^{n} \theta_{S(i)T(j)} A_{ij}$$

$$\to \langle \theta, A/\hat{S}\hat{T} \rangle, \text{ where } \hat{S} \text{ and } \hat{T} \text{ given by:}$$



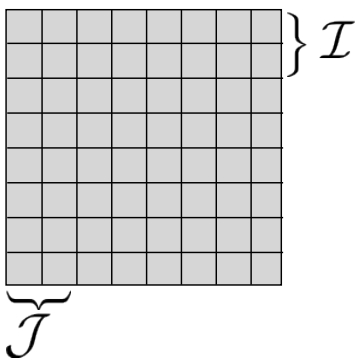
Picture of near-optimal clustering

Similar to Borgs et al. (2008), Alon et al. (2003)

$$\max_{S,T} \langle \theta, A/ST \rangle = \max_{ST} \sum_{i=1}^{m} \sum_{j=1}^{n} \theta_{S(i)T(j)} A_{ij}$$

$$\rightarrow \langle \theta, A/\hat{S}\hat{T} \rangle, \text{ where } \hat{S} \text{ and } \hat{T} \text{ given by:}$$

- Randomly sample rows $\mathcal{I}$ and columns $\mathcal{J}$

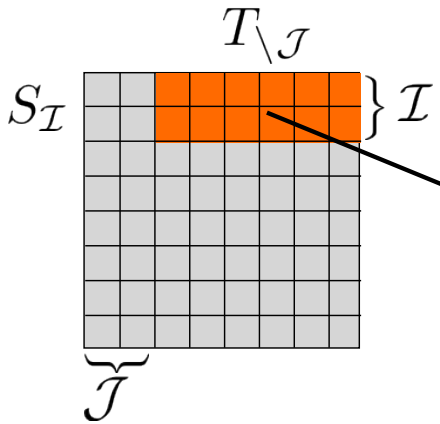


Picture of near-optimal clustering

Similar to Borgs et al. (2008), Alon et al. (2003)

$$\max_{S,T} \langle \theta, A/ST \rangle = \max_{ST} \sum_{i=1}^{m} \sum_{j=1}^{n} \theta_{S(i)T(j)} A_{ij}$$

$$\to \langle \theta, A/\hat{S}\hat{T} \rangle, \text{ where } \hat{S} \text{ and } \hat{T} \text{ given by:}$$



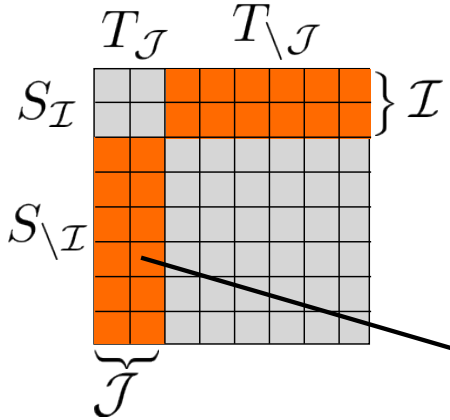
- Randomly sample rows $\mathcal{I}$ and columns $\mathcal{J}$
- Given $S_{\mathcal{I}}$, choose $T_{\setminus \mathcal{J}}$ to maximize $\sum_{j \notin \mathcal{J}} \sum_{i \in \mathcal{I}} \theta_{S(i)T(j)} \omega(\xi_i, \zeta_j)$.

Picture of near-optimal clustering

Similar to Borgs et al. (2008), Alon et al. (2003)

$$\max_{S,T} \langle \theta, A/ST \rangle = \max_{ST} \sum_{i=1}^{m} \sum_{j=1}^{n} \theta_{S(i)T(j)} A_{ij}$$

$$\to \langle \theta, A/\hat{S}\hat{T} \rangle, \text{ where } \hat{S} \text{ and } \hat{T} \text{ given by:}$$



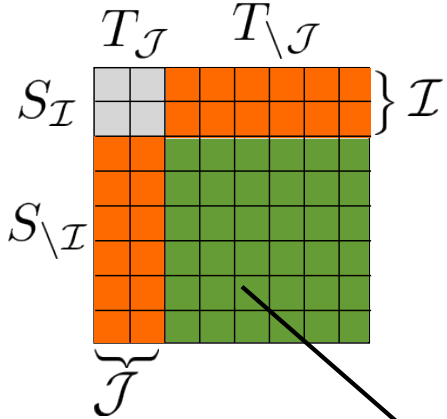
- Randomly sample rows $\mathcal{I}$ and columns $\mathcal{J}$
- Given $S_{\mathcal{I}}$, choose $T_{\setminus\mathcal{J}}$ to maximize $\sum_{j \notin \mathcal{J}} \sum_{i \in \mathcal{I}} \theta_{S(i)T(j)} \omega(\xi_i, \zeta_j)$.
- Given $T_{\mathcal{J}}$, choose $S_{\setminus\mathcal{I}}$ to maximize $\sum_{i \notin \mathcal{I}} \sum_{j \in \mathcal{J}} \theta_{S(i)T(j)} \omega(\xi_i, \zeta_j)$.

Picture of near-optimal clustering

Similar to Borgs et al. (2008), Alon et al. (2003)

$$\max_{S,T} \langle \theta, A/ST \rangle = \max_{ST} \sum_{i=1}^{m} \sum_{j=1}^{n} \theta_{S(i)T(j)} A_{ij}$$

$$\rightarrow \langle \theta, A/\hat{S}\hat{T} \rangle, \text{ where } \hat{S} \text{ and } \hat{T} \text{ given by:}$$

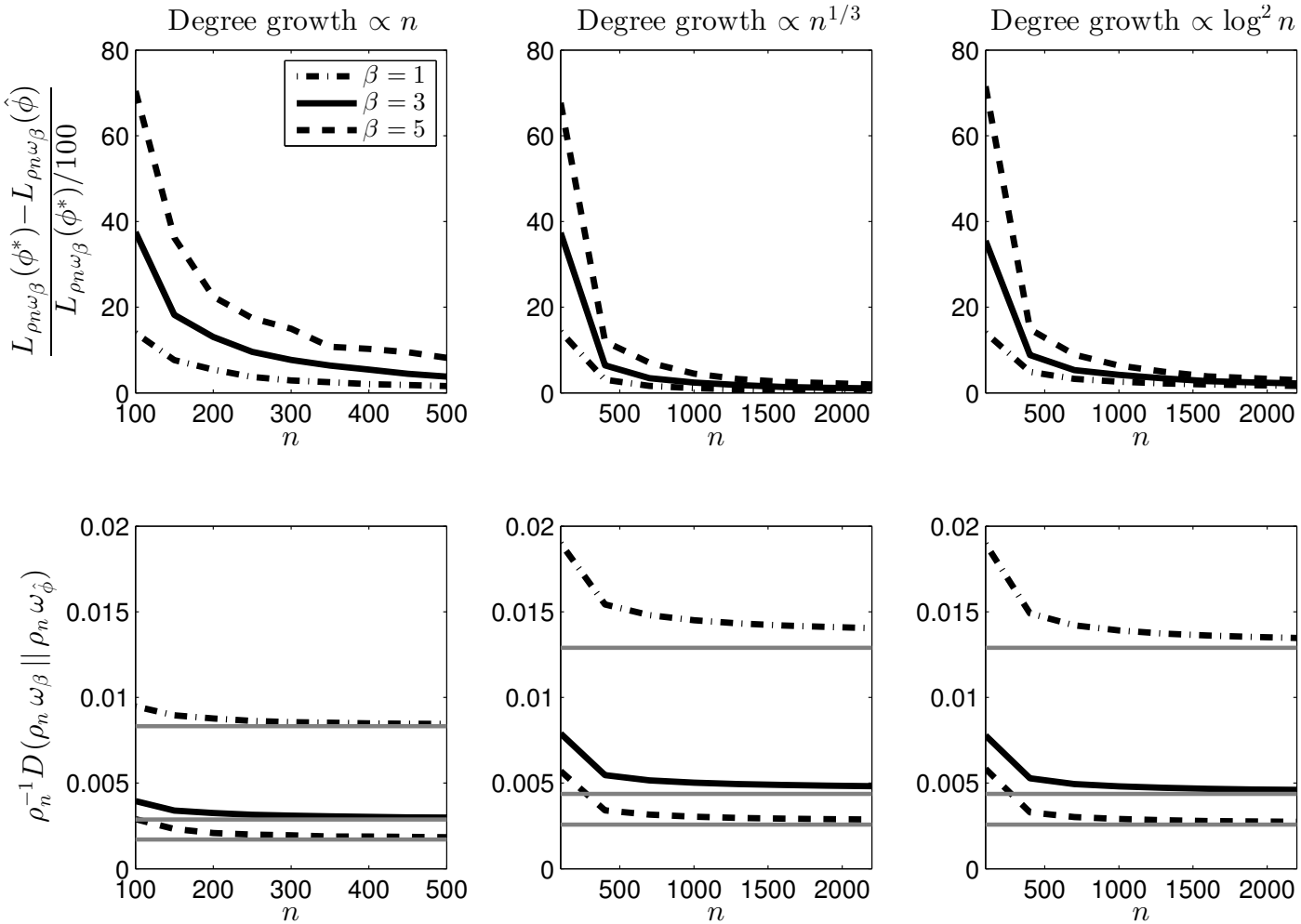


- Randomly sample rows $\mathcal{I}$ and columns $\mathcal{J}$
- Given $S_{\mathcal{I}}$, choose $T_{\setminus\mathcal{J}}$ to maximize $\sum_{j \notin \mathcal{J}} \sum_{i \in \mathcal{I}} \theta_{S(i)T(j)} \omega(\xi_i, \zeta_j)$.
- Given $T_{\mathcal{J}}$, choose $S_{\setminus\mathcal{I}}$ to maximize $\sum_{i \notin \mathcal{I}} \sum_{j \in \mathcal{J}} \theta_{S(i)T(j)} \omega(\xi_i, \zeta_j)$.
- Choose $S_{\mathcal{I}}$ and $T_{\mathcal{J}}$ to maximize $\sum_{i \notin \mathcal{I}} \sum_{j \notin \mathcal{J}} \theta_{S(i)T(j)} \omega(\xi_i \zeta_j)$

Show that $\hat{S}$ and $\hat{T}$ involve $K^{|\mathcal{I}|+|\mathcal{J}|} \ll K^n$ possible labelings

Simulation results

Data generated by non-blockmodel



Top: excess risk.

Bottom: KL divergence of blockmodel approximation (grey line is optimal).

β: True model more resembles blockmodel for large β.