A Tensor Spectral Approach to Learning Mixed Membership Community Models

Anima Anandkumar

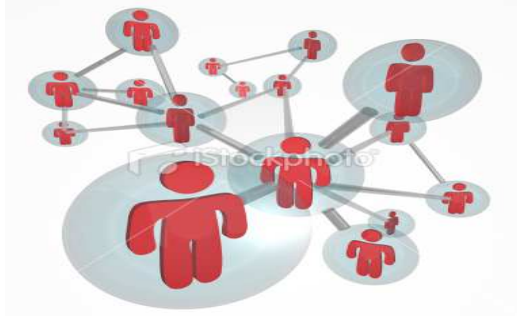
U.C. Irvine

Joint work with Rong Ge, Daniel Hsu, Furong Huang,
Niranjan UN, Mohammad Hakeem, Sham Kakade.

Network Communities in Various Domains

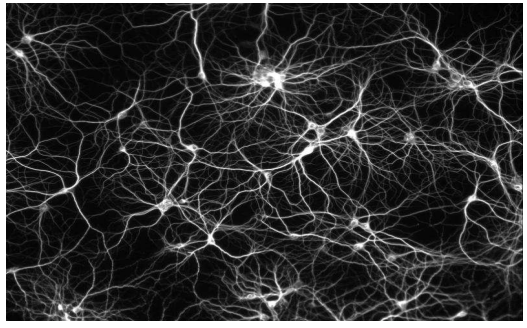
Social Networks

- Social ties: e.g. friendships, co-authorships



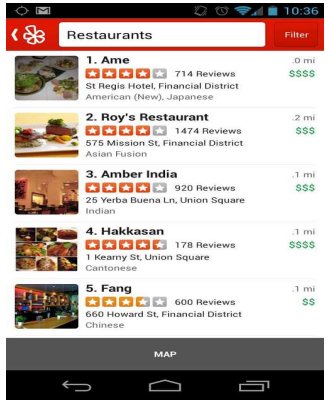
Biological Networks

- Functional relationships:
  e.g. gene regulation, neural activity.



Recommendation Systems

- Recommendations: e.g. yelp reviews.



Community Detection: Infer hidden communities from observed network.

Community Formation Models

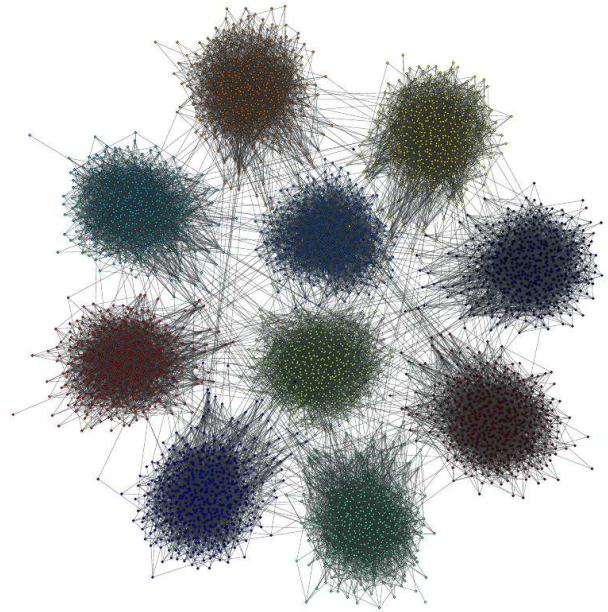
Basic Intuition: Nodes connect due to their community memberships

Community Formation Models

Basic Intuition: Nodes connect due to their community memberships

Classical: Stochastic Block Model

- Edges conditionally independent given node community memberships
- Single membership model: Nodes in at most one community

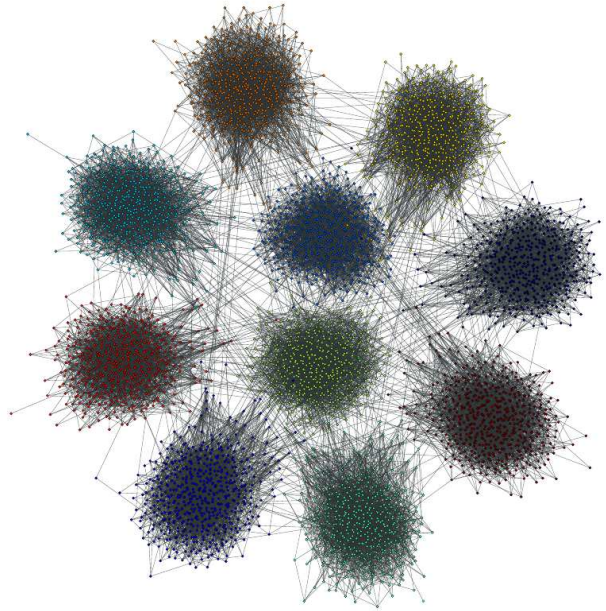


Community Formation Models

Basic Intuition: Nodes connect due to their community memberships

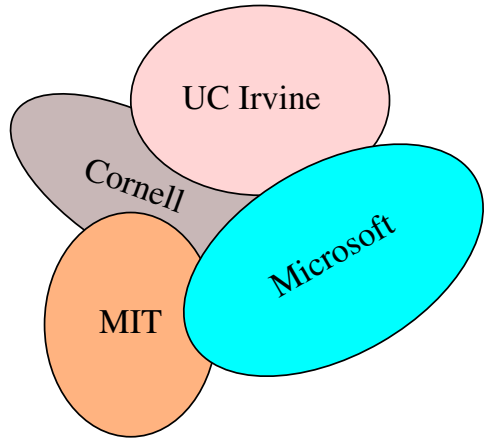
Classical: Stochastic Block Model

- Edges conditionally independent given node community memberships
- Single membership model: Nodes in at most one community



Modeling Overlapping Communities

- People belong to multiple communities
- Community formation models?
- Detection algorithms?
  Computational/sample complexities?



Mixed Membership Community Models

Node Membership Model

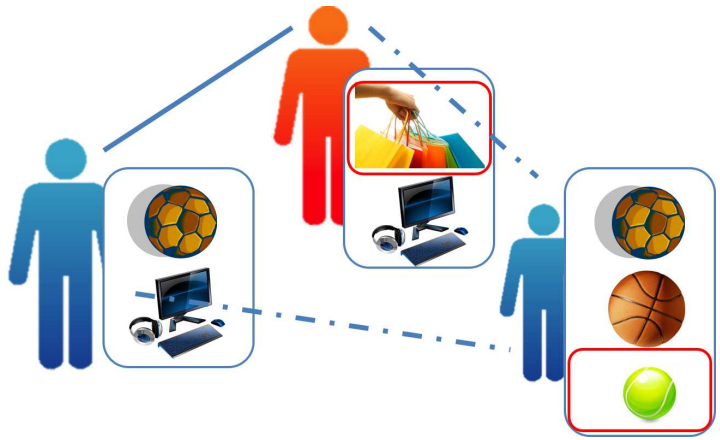
- Mixed memberships: Nodes can belong to multiple communities
- Fractional memberships: Node memberships normalized to one.

Mixed Membership Community Models

Node Membership Model

- Mixed memberships: Nodes can belong to multiple communities
- Fractional memberships: Node memberships normalized to one.

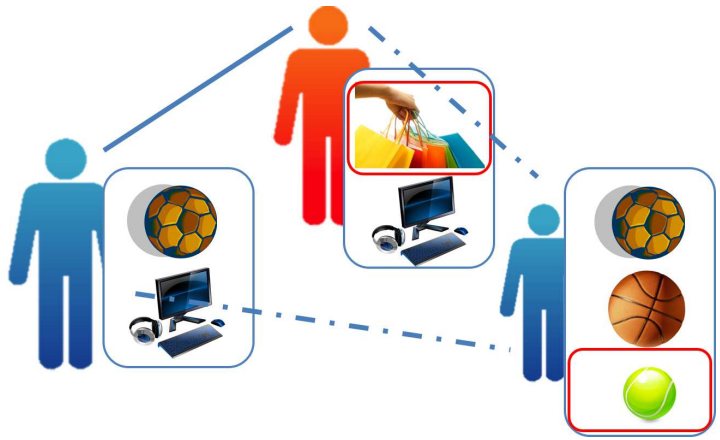
Mixed Membership Community Models



Node Membership Model

- Mixed memberships: Nodes can belong to multiple communities
- Fractional memberships: Node memberships normalized to one.

Mixed Membership Community Models



Node Membership Model

- Mixed memberships: Nodes can belong to multiple communities
- Fractional memberships: Node memberships normalized to one.

Edge Formation Model

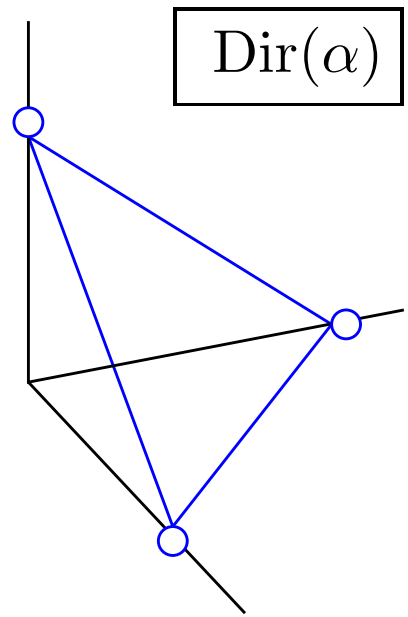
- Edges conditionally independent given node community memberships
- Linearity: Edge probability averaged over community memberships

Mixed Membership Dirichlet Model (Airoldi et. al.)

- Independent draws for community membership vectors $\{\pi_u\}_{u \in V}$ from Dirichlet distribution

$$\boxed{\mathbb{P}[\pi_u] \propto \prod_{j=1}^{k} \pi_u(j)^{\alpha_j - 1}, \qquad \sum_{j=1}^{k} \pi_u(j) = 1}$$

- Dirichlet distribution supported over simplex



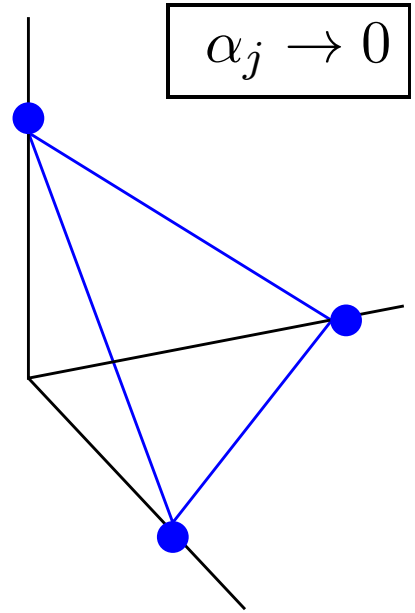
Mixed Membership Dirichlet Model (Airoldi et. al.)

- Independent draws for community membership vectors $\{\pi_u\}_{u \in V}$ from Dirichlet distribution

$$\boxed{\mathbb{P}[\pi_u] \propto \prod_{j=1}^{k} \pi_u(j)^{\alpha_j - 1}, \qquad \sum_{j=1}^{k} \pi_u(j) = 1}$$

- Dirichlet distribution supported over simplex

$\boxed{\alpha_j \to 0}$



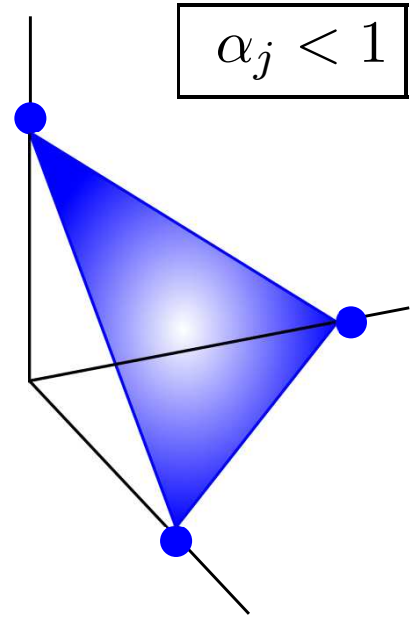
Mixed Membership Dirichlet Model (Airoldi et. al.)

- Independent draws for community membership vectors $\{\pi_u\}_{u \in V}$ from Dirichlet distribution

$$\mathbb{P}[\pi_u] \propto \prod_{j=1}^{k} \pi_u(j)^{\alpha_j - 1}, \qquad \sum_{j=1}^{k} \pi_u(j) = 1$$

- Dirichlet distribution supported over simplex

$\alpha_j < 1$



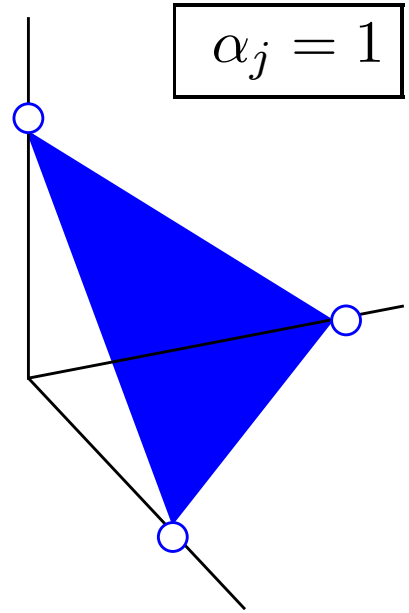
Mixed Membership Dirichlet Model (Airoldi et. al.)

- Independent draws for community membership vectors $\{\pi_u\}_{u \in V}$ from Dirichlet distribution

$$\mathbb{P}[\pi_u] \propto \prod_{j=1}^{k} \pi_u(j)^{\alpha_j - 1}, \qquad \sum_{j=1}^{k} \pi_u(j) = 1$$

- Dirichlet distribution supported over simplex

$\alpha_j = 1$

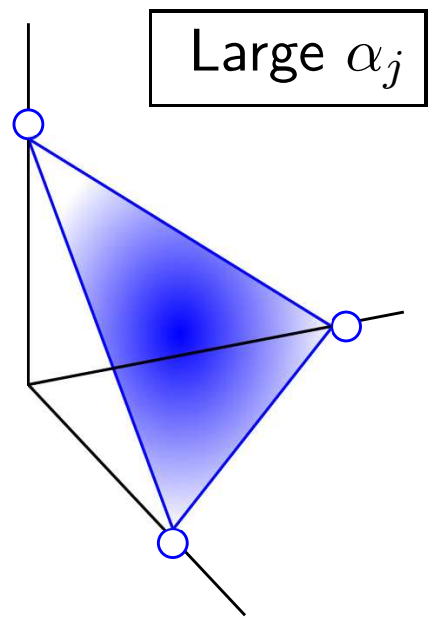


Mixed Membership Dirichlet Model (Airoldi et. al.)

- Independent draws for community membership vectors $\{\pi_u\}_{u \in V}$ from Dirichlet distribution

$$\mathbb{P}[\pi_u] \propto \prod_{j=1}^{k} \pi_u(j)^{\alpha_j - 1}, \qquad \sum_{j=1}^{k} \pi_u(j) = 1$$

- Dirichlet distribution supported over simplex



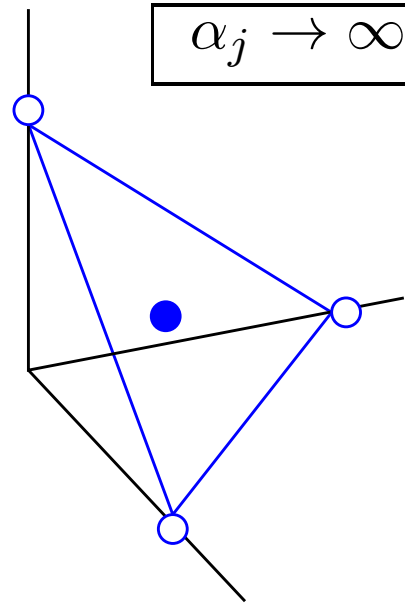
Mixed Membership Dirichlet Model (Airoldi et. al.)

- Independent draws for community membership vectors $\{\pi_u\}_{u \in V}$ from Dirichlet distribution

$$\boxed{\mathbb{P}[\pi_u] \propto \prod_{j=1}^{k} \pi_u(j)^{\alpha_j - 1}, \qquad \sum_{j=1}^{k} \pi_u(j) = 1}$$

- Dirichlet distribution supported over simplex

$\boxed{\alpha_j \to \infty}$

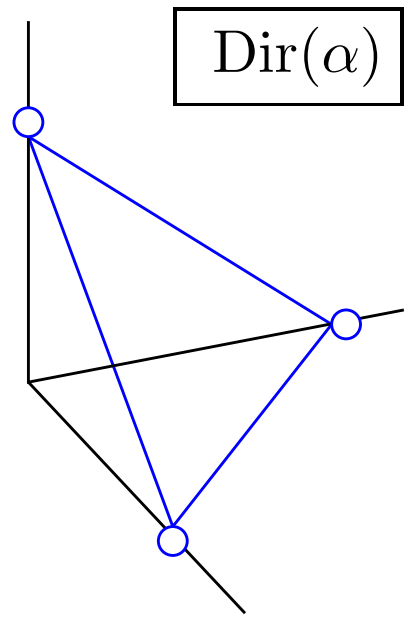


Mixed Membership Dirichlet Model (Airoldi et. al.)

- Independent draws for community membership vectors $\{\pi_u\}_{u \in V}$ from Dirichlet distribution

$$\boxed{\mathbb{P}[\pi_u] \propto \prod_{j=1}^{k} \pi_u(j)^{\alpha_j - 1}, \qquad \sum_{j=1}^{k} \pi_u(j) = 1}$$

- Dirichlet distribution supported over simplex



Mixed Membership Dirichlet Model (Airoldi et. al.)

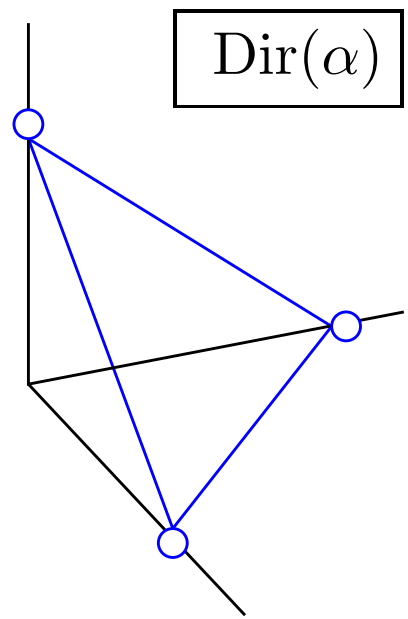
- Independent draws for community membership vectors $\{\pi_u\}_{u \in V}$ from Dirichlet distribution

$$\boxed{\mathbb{P}[\pi_u] \propto \prod_{j=1}^{k} \pi_u(j)^{\alpha_j - 1}, \qquad \sum_{j=1}^{k} \pi_u(j) = 1}$$

- Dirichlet distribution supported over simplex
- Dirichlet concentration parameter

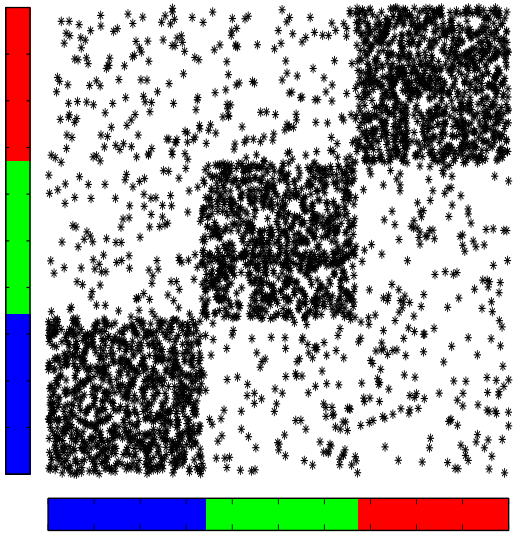
$$\boxed{\alpha_0 := \sum_j \alpha_j}$$

- Sparsity level in π_u is $O(\alpha_0)$.
- Regime of interest: small α_0



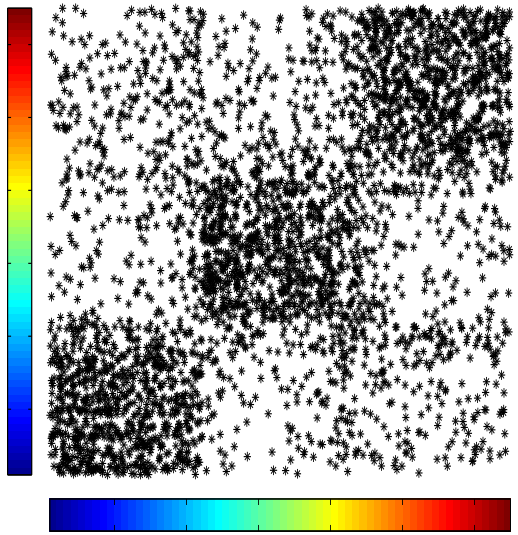
Pure vs. Mixed Membership Community Models

Stochastic Block Model



$\alpha_0 = 0$

Mixed Membership Model



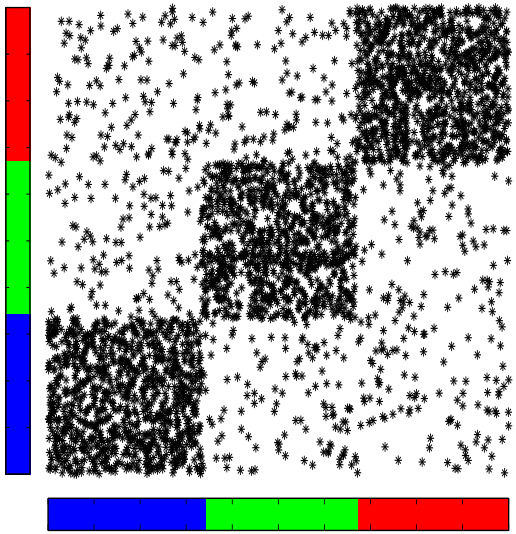
$\alpha_0 = 1$

Challenges in Learning Mixed Membership Models

- Identifiability: when can parameters be estimated?
- Guaranteed learning? What input required?
- Potentially large sample and computational complexities

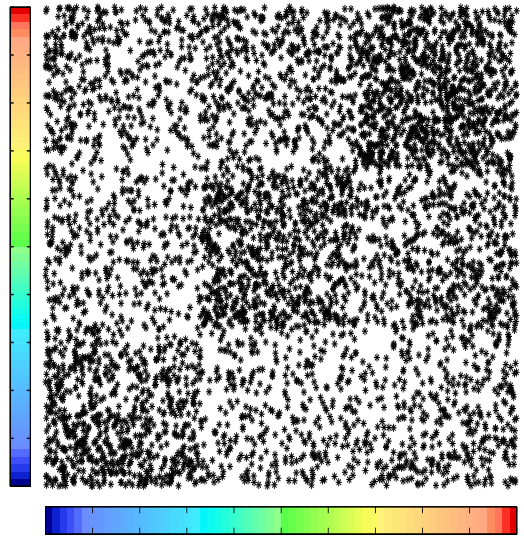
Pure vs. Mixed Membership Community Models

Stochastic Block Model



$\alpha_0 = 0$

Mixed Membership Model



$\alpha_0 = 10$

Challenges in Learning Mixed Membership Models

- Identifiability: when can parameters be estimated?
- Guaranteed learning? What input required?
- Potentially large sample and computational complexities

Outline

Summary of Results

Contributions

- First guaranteed learning method for overlapping (probabilistic) community models.
- Correctness under exact moments: edges and 3-star counts.
- Efficient sample and computational complexity.

Summary of Results

Contributions

- First guaranteed learning method for overlapping (probabilistic) community models.
- Correctness under exact moments: edges and 3-star counts.
- Efficient sample and computational complexity.

Scaling Requirements

k communities, n nodes. Uniform communities. Dirichlet parameter: $\alpha_0 := \sum_i \alpha_i$. p, q: intra/inter-community connectivity

$$n = \tilde{\Omega}(k^2(\alpha_0+1)^2), \qquad \frac{p-q}{\sqrt{p}} = \tilde{\Omega}\left(\frac{(\alpha_0+1)k}{\sqrt{n}}\right).$$

Summary of Results

Contributions

- First guaranteed learning method for overlapping (probabilistic) community models.
- Correctness under exact moments: edges and 3-star counts.
- Efficient sample and computational complexity.

Scaling Requirements

k communities, n nodes. Uniform communities. Dirichlet parameter: $\alpha_0 := \sum_i \alpha_i$. p, q: intra/inter-community connectivity

$$n = \tilde{\Omega}(k^2(\alpha_0+1)^2), \qquad \frac{p-q}{\sqrt{p}} = \tilde{\Omega}\left(\frac{(\alpha_0+1)k}{\sqrt{n}}\right).$$

- For stochastic block model $(\alpha_0 = 0)$, tight results
- Performance degradation as α_0 increases
- Efficient method for sparse community overlaps

Main Results: Recovery Guarantees

- k communities, n nodes. Uniform communities.
- Community membership matrix Π, $\Pi^{(i)}$: i^{th} community
- Connectivity matrix P: $P(i,i) = p$ and $P(i,j) = q$ for $i \neq j$.

Scaling Requirements

$$n = \tilde{\Omega}(k^2(\alpha_0+1)^2), \qquad \frac{p-q}{\sqrt{p}} = \tilde{\Omega}\left(\frac{(\alpha_0+1)k}{\sqrt{n}}\right).$$

Main Results: Recovery Guarantees

- k communities, n nodes. Uniform communities.
- Community membership matrix Π, $\Pi^{(i)}$: i^{th} community
- Connectivity matrix P: $P(i,i) = p$ and $P(i,j) = q$ for $i \neq j$.

Scaling Requirements

$$n = \tilde{\Omega}(k^2(\alpha_0+1)^2), \qquad \frac{p-q}{\sqrt{p}} = \tilde{\Omega}\left(\frac{(\alpha_0+1)k}{\sqrt{n}}\right).$$

Recovery Bounds (Anandkumar, Ge, Hsu, Kakade '13)

$$\frac{\varepsilon_\pi}{n} := \frac{1}{n}\max_i \|\widehat{\Pi}^i - \Pi^i\|_1 = \tilde{O}\left(\frac{(\alpha_0+1)^{3/2}\sqrt{p}}{(p-q)\sqrt{n}}\right)$$

$$\varepsilon_P := \max_{i,j\in[n]} |\widehat{P}_{i,j} - P_{i,j}| = \tilde{O}\left(\frac{(\alpha_0+1)^{3/2}k\sqrt{p}}{(p-q)\sqrt{n}}\right)$$

Support Recovery Guarantees (Homophilic Models)

- ε_P : Error in recovering P
- Π: true community membership matrix.
- Homophilic Models: $p > q$
- $\widehat{S}$: Estimated supports.

Support Recovery Guarantees (Homophilic Models)

- ε_P : Error in recovering P
- Π: true community membership matrix.
- Homophilic Models: $p > q$
- $\widehat{S}$: Estimated supports.

Support Recovery Guarantee (AGHK '13)

For a threshold $\boxed{\xi = \Omega(\varepsilon_P)}$, for all nodes $j \in [n]$ and all communities $i \in [k]$, the estimated support $\widehat{S}$ satisfies (w.h.p)

$$\Pi(i,j) \geq \xi \Rightarrow \widehat{S}(i,j) = 1 \quad \text{and} \quad \Pi(i,j) \leq \frac{\xi}{2} \Rightarrow \widehat{S}(i,j) = 0.$$

Zero-error Support Recovery of Significant Memberships of All Nodes

Efficient Recovery of Mixed Memberships

Overview of the Approach

- Inverse moment method: fit parameters to observed moments
- Tensor spectral approach: Compute "spectrum" of tensor computed from moments
- Non-convex but computationally tractable approaches

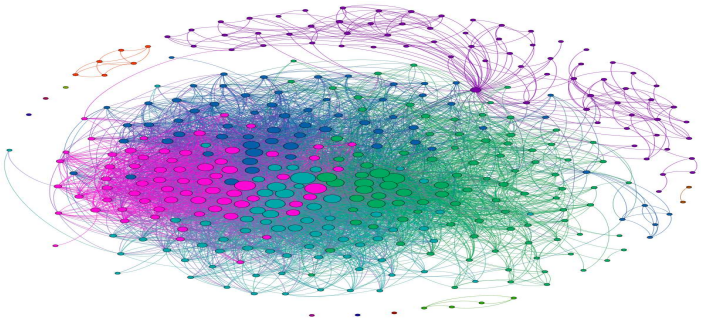
Overview of the Approach

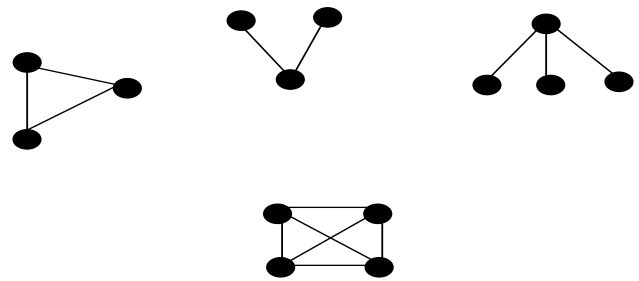
- Inverse moment method: fit parameters to observed moments
- Tensor spectral approach: Compute "spectrum" of tensor computed from moments
- Non-convex but computationally tractable approaches

- Inverse moment method: subgraph counts
- Tensor spectral approach: Low rank tensor form and efficient decomposition via power method
- Efficient Implementation: Linear algebraic operations and online tensor decomposition

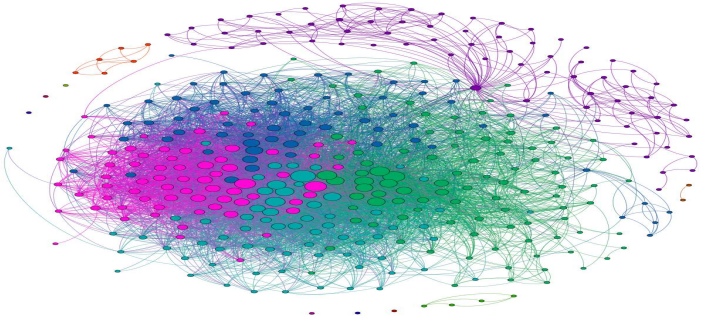
Outline

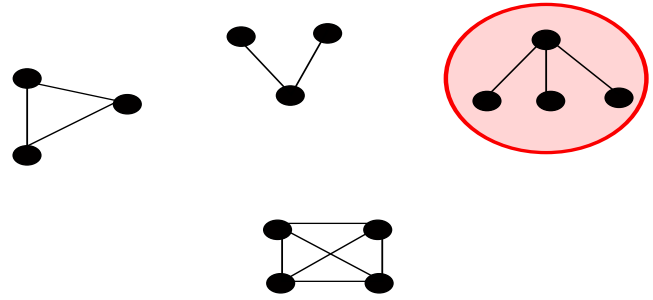
Subgraph Counts as Graph Moments



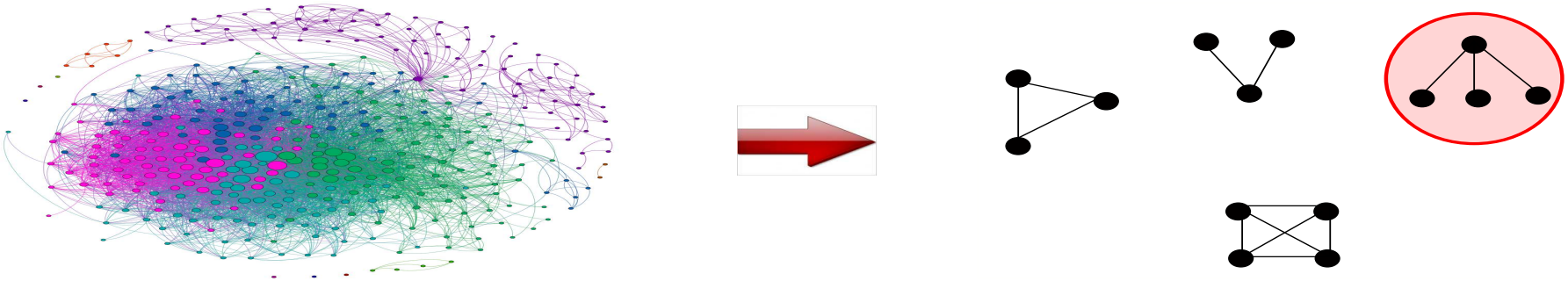


Subgraph Counts as Graph Moments



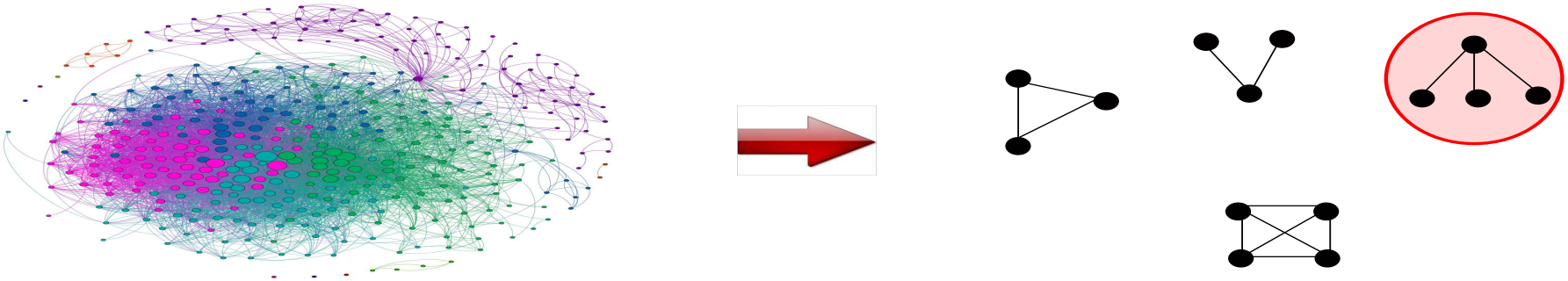


Subgraph Counts as Graph Moments



3-star counts sufficient for identifiability and learning of MMSB

Subgraph Counts as Graph Moments



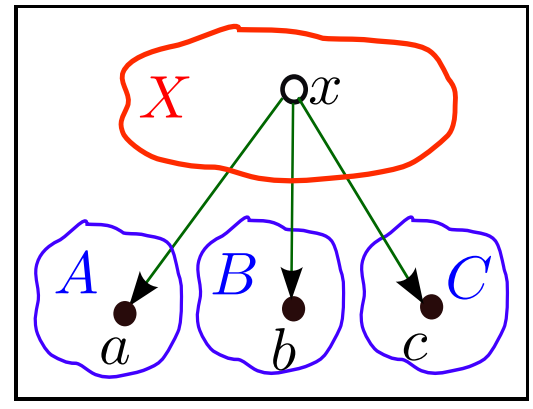
3-star counts sufficient for identifiability and learning of MMSB

3-Star Count Tensor

$$M_3(a,b,c) = \frac{1}{|X|} \# \text{ of 3-stars with leaves a,b,c}$$

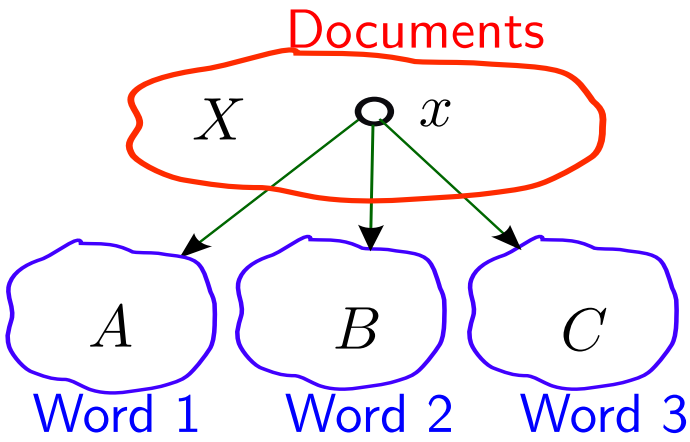
$$= \frac{1}{|X|} \sum_{x \in X} G(x,a)G(x,b)G(x,c).$$

$$M_3 = \frac{1}{|X|} \sum_{x \in X} [G_{x,A}^\top \otimes G_{x,B}^\top \otimes G_{x,C}^\top]$$

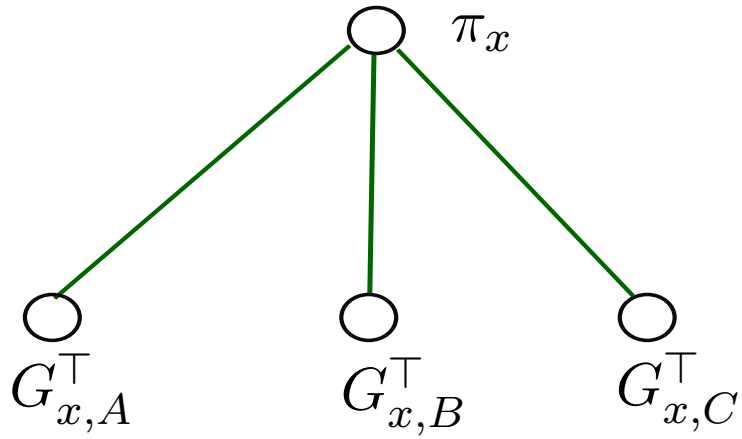


Topic Model or Multi-view Representation

Topic model



Graphical model



- Conditional independence of the three words or views
- Exploit to find expected M_3

$$M_3 = \frac{1}{|X|} \sum_{x \in X} [G_{x,A}^\top \otimes G_{x,B}^\top \otimes G_{x,C}^\top]$$

"Tensor Decompositions for Learning Latent Variable Models" by A. Anandkumar, R. Ge, D. Hsu, S.M. Kakade and M. Telgarsky. Preprint, October 2012.

Moments under Stochastic Block Model

- One-hot encoding: $\boxed{\pi_u = e_i}$ if node u is in community i.
- $\lambda_i = \mathbb{P}[\pi = e_i]$: probability of community i.
- $\boxed{\mathbb{P}[G_{u,v} = 1 | \pi_u, \pi_v] = \pi_u^\top P \pi_v}$. E.g. $\pi_u = e_i$, $\pi_v = e_j$, prob. is $P_{i,j}$.

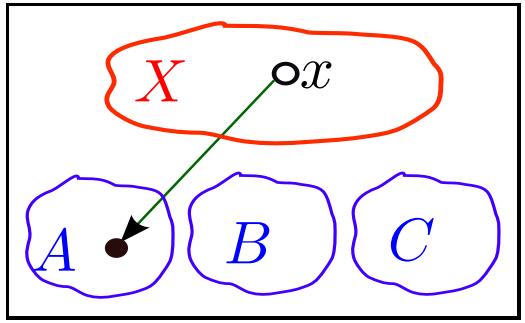
Moments under Stochastic Block Model

- One-hot encoding: $\boxed{\pi_u = e_i}$ if node u is in community i.
- $\lambda_i = \mathbb{P}[\pi = e_i]$: probability of community i.
- $\boxed{\mathbb{P}[G_{u,v} = 1 | \pi_u, \pi_v] = \pi_u^\top P \pi_v}$. E.g. $\pi_u = e_i$, $\pi_v = e_j$, prob. is $P_{i,j}$.

Expected Edge Counts

- Community matrix: $\Pi_A := [\pi_a]_{a \in A}$

$$\boxed{\mathbb{E}[G_{x,A}^\top | \Pi] = \pi_x^\top P \Pi_A = \Pi_A^\top P^\top \pi_x = F_A \pi_x}$$



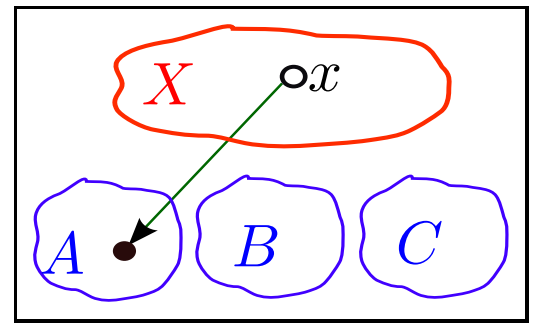
Moments under Stochastic Block Model

- One-hot encoding: $\boxed{\pi_u = e_i}$ if node u is in community i.
- $\lambda_i = \mathbb{P}[\pi = e_i]$: probability of community i.
- $\boxed{\mathbb{P}[G_{u,v} = 1 | \pi_u, \pi_v] = \pi_u^\top P \pi_v}$. E.g. $\pi_u = e_i$, $\pi_v = e_j$, prob. is $P_{i,j}$.

Expected Edge Counts

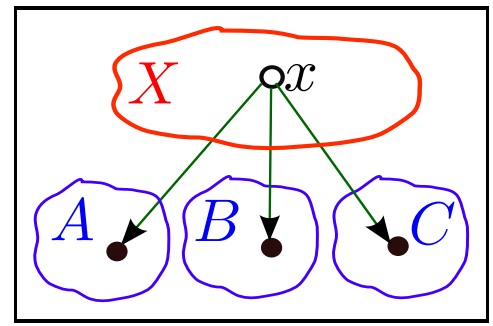
- Community matrix: $\Pi_A := [\pi_a]_{a \in A}$

$$\boxed{\mathbb{E}[G_{x,A}^\top | \Pi] = \pi_x^\top P \Pi_A = \Pi_A^\top P^\top \pi_x = F_A \pi_x}$$



Expected 3-Star Tensor

$\mathbb{E}[M_3 | \Pi_{A,B,C}] = ?$



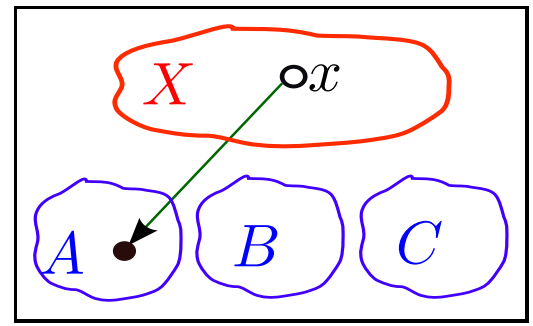
Moments under Stochastic Block Model

- One-hot encoding: $\boxed{\pi_u = e_i}$ if node u is in community i.
- $\lambda_i = \mathbb{P}[\pi = e_i]$: probability of community i.
- $\boxed{\mathbb{P}[G_{u,v} = 1 | \pi_u, \pi_v] = \pi_u^\top P \pi_v}$. E.g. $\pi_u = e_i$, $\pi_v = e_j$, prob. is $P_{i,j}$.

Expected Edge Counts

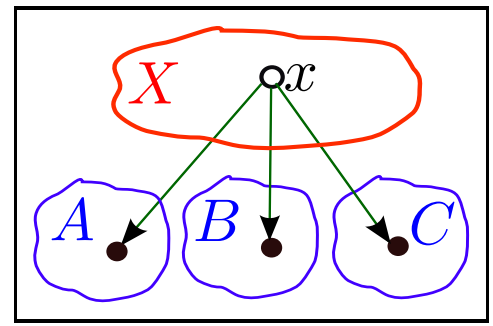
- Community matrix: $\Pi_A := [\pi_a]_{a \in A}$

$$\boxed{\mathbb{E}[G_{x,A}^\top | \Pi] = \pi_x^\top P \Pi_A = \Pi_A^\top P^\top \pi_x = F_A \pi_x}$$



Expected 3-Star Tensor

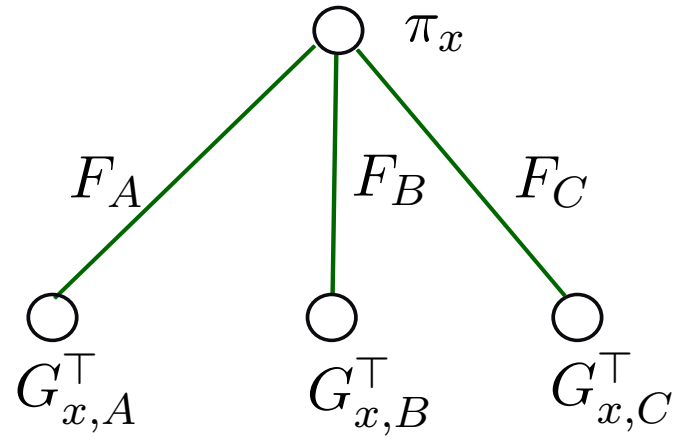
$\mathbb{E}[M_3 | \Pi_{A,B,C}] = ?$



3-star Tensor Form through Multi-view Model

3-Star Tensor Form

$$M_3 = \frac{1}{|X|} \sum_{x \in X} [G_{x,A}^\top \otimes G_{x,B}^\top \otimes G_{x,C}^\top]$$



Linear Multiview Model: $\mathbb{E}[G_{x,A}^\top | \Pi] = F_A \pi_x$, Independent views

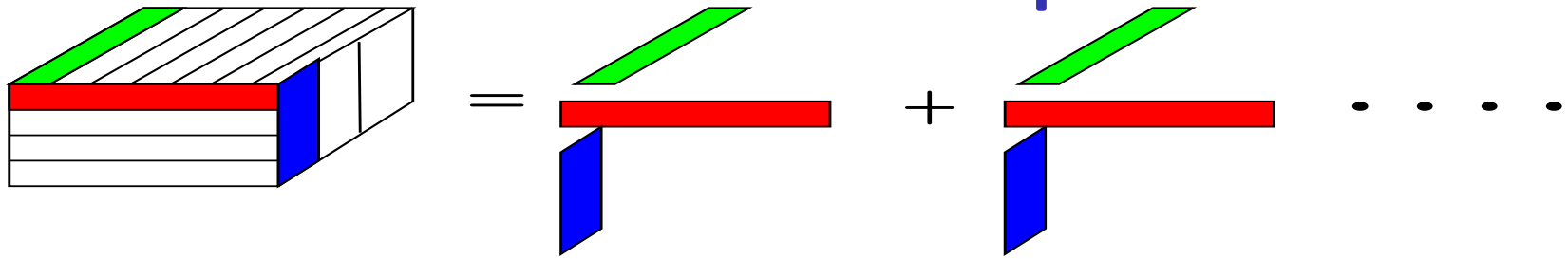
$$\mathbb{E}[M_3 | \Pi_{A,B,C,X}] = \sum_{x \in X} \frac{1}{|X|} [(F_A \pi_x) \otimes (F_B \pi_x) \otimes (F_C \pi_x)]$$

$$\mathbb{E}[M_3 | \Pi_{A,B,C}] = \sum_{i \in [k]} \lambda_i [(F_A)_i \otimes (F_B)_i \otimes (F_C)_i]$$

Goal: Recover $F_A, F_B, F_C, \vec{\lambda}$ through CP tensor decomposition

Outline

Low-rank Tensor Decomposition

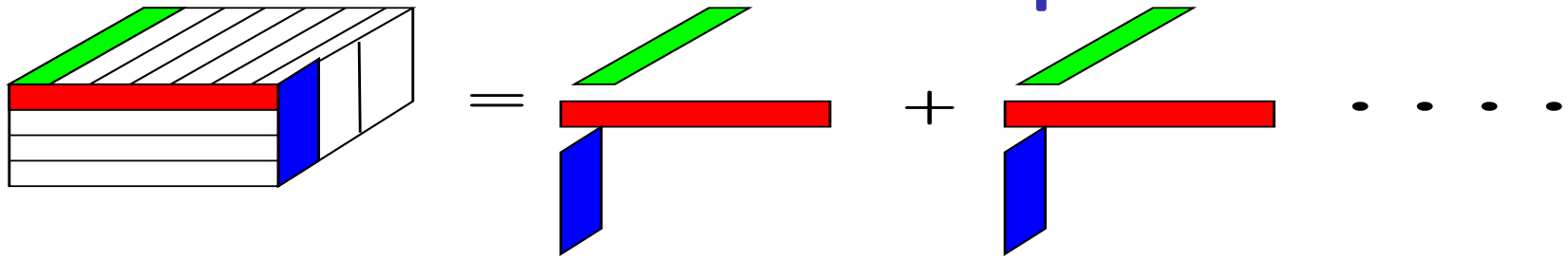


Tensor $\mathbb{E}[M_3|\Pi_{A,B,C}]$ $\lambda_1(F_A)_1 \otimes (F_B)_1 \otimes (F_C)_1$ $\lambda_2(F_A)_2 \otimes (F_B)_2 \otimes (F_C)_2$

$$\mathbb{E}[M_3|\Pi_{A,B,C}] = \sum_{i\in[k]} \lambda_i[(F_A)_i \otimes (F_B)_i \otimes (F_C)_i]$$

- Rank-k tensor decomposition and typically $k \ll n$
- $u \otimes v \otimes w$ is a rank-1 tensor whose i, j, k^{th} entry is $u_i v_j w_k$.

Low-rank Tensor Decomposition



Tensor $\mathbb{E}[M_3|\Pi_{A,B,C}]$ $\lambda_1(F_A)_1 \otimes (F_B)_1 \otimes (F_C)_1$ $\lambda_2(F_A)_2 \otimes (F_B)_2 \otimes (F_C)_2$

$$\mathbb{E}[M_3|\Pi_{A,B,C}] = \sum_{i\in[k]} \lambda_i[(F_A)_i \otimes (F_B)_i \otimes (F_C)_i]$$

- Rank-k tensor decomposition and typically $k \ll n$
- $u \otimes v \otimes w$ is a rank-1 tensor whose i, j, k^{th} entry is $u_i v_j w_k$.

Challenges

- Guaranteed algorithm for tensor decomposition?
- Efficient and scalable implementation?
- Noisy tensor decomposition: exact moments not available
- Sample complexity? How large n compared to k?

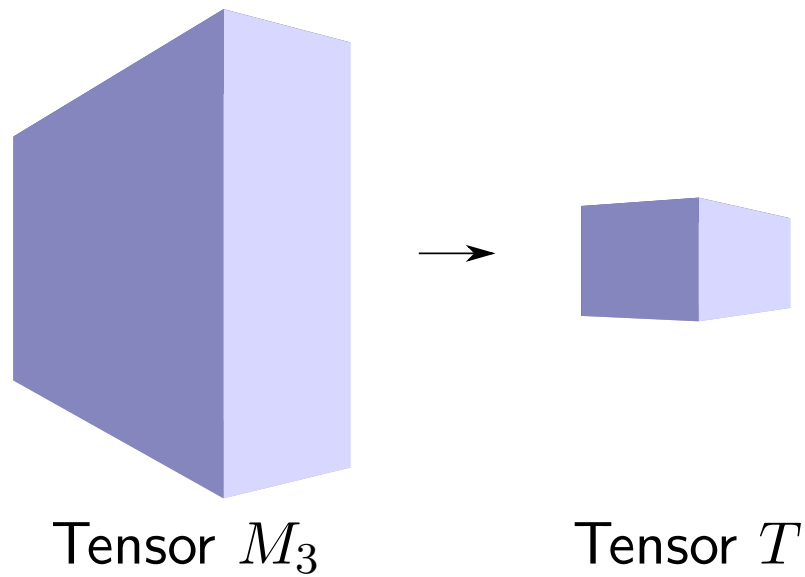
Dimensionality Reduction for Tensor Decomposition

$$\mathbb{E}[M_3|\Pi_{A,B,C}] = \sum_{i\in[k]} \lambda_i[(F_A)_i \otimes (F_B)_i \otimes (F_C)_i]$$

- Rank-k tensor decomposition and typically $k \ll n$
- M_3 has size $O(n^3)$ but number of free parameters: $nk + k$

First Step: Dimensionality Reduction

- Convert M_3 of size $|A| \times |B| \times |C|$ to a tensor T of size $k \times k \times k$
- Carry out decomposition of T



Tensor M_3 Tensor T

Advantages

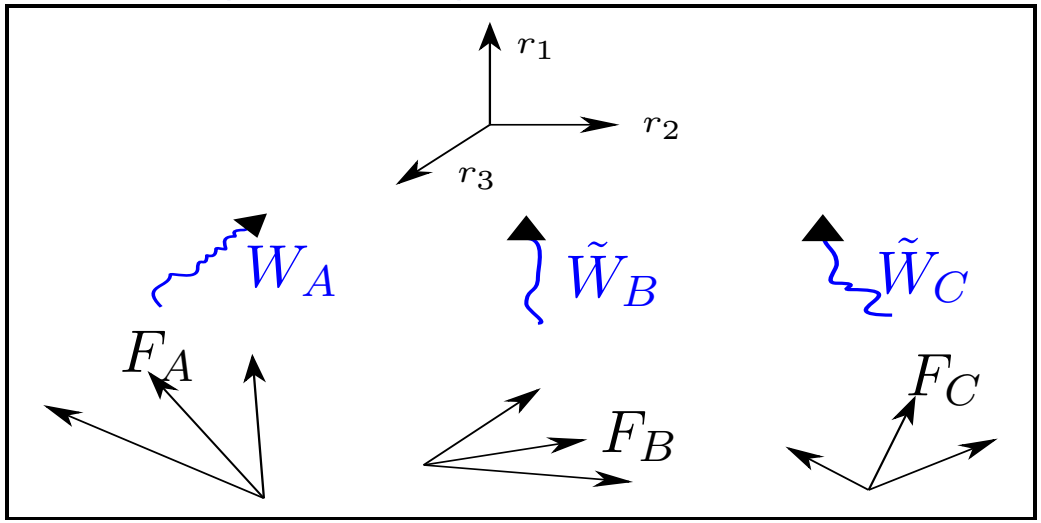
- Reduced computation
- Robustness to noise

Dimensionality reduction through multi-linear transforms

Dimensionality Reduction through Whitening

$$\boxed{\mathbb{E}[M_3|\Pi_{A,B,C}] = \sum_i \lambda_i [(F_A)_i \otimes (F_B)_i \otimes (F_C)_i]}$$

Whitening: Conversion of (expected) M_3 to Orthogonal Symmetric Tensor T



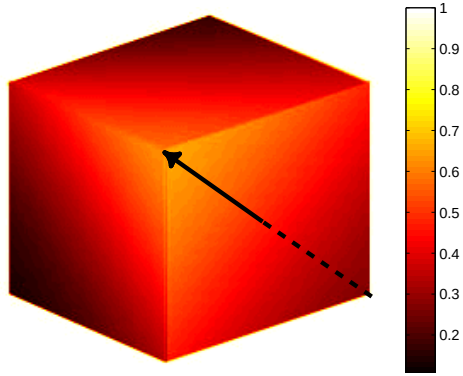
Multi-linear Transformation of 3-star Tensor

$$\boxed{T := \mathbb{E}[M_3|\Pi_{A,B,C}](W_A, \tilde{W}_B, \tilde{W}_C) = \sum_i \rho_i r_i^{\otimes 3}}$$

- T is symmetric orthogonal tensor: $\{r_i\}$ are orthonormal.

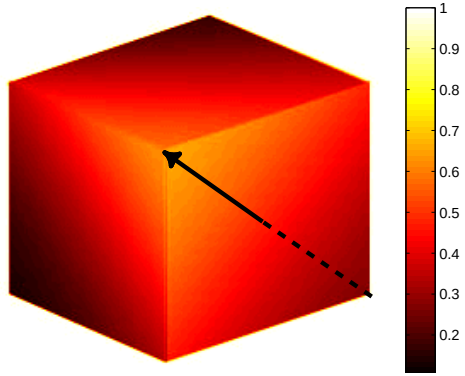
Tensor Decomposition Through Eigen Analysis

- Orthogonal symmetric tensor: $T = \sum_i \rho_i r_i^{\otimes 3}$
- $T(I, r_i, r_i) = \sum_j \rho_j \langle r_i, r_j \rangle^2 r_j = \rho_i r_i$



Tensor Decomposition Through Eigen Analysis

- Orthogonal symmetric tensor: $T = \sum_i \rho_i r_i^{\otimes 3}$
- $T(I, r_i, r_i) = \sum_j \rho_j \langle r_i, r_j \rangle^2 r_j = \rho_i r_i$

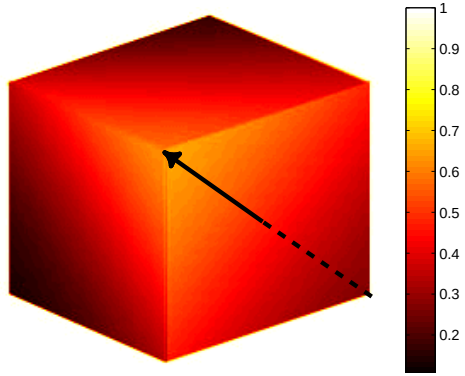


Obtaining eigenvectors through power iterations

$$u \mapsto \frac{T(I, u, u)}{\|T(I, u, u)\|}$$

Tensor Decomposition Through Eigen Analysis

- Orthogonal symmetric tensor: $T = \sum_i \rho_i r_i^{\otimes 3}$
- $T(I, r_i, r_i) = \sum_j \rho_j \langle r_i, r_j \rangle^2 r_j = \rho_i r_i$



Obtaining eigenvectors through power iterations

$$u \mapsto \frac{T(I, u, u)}{\|T(I, u, u)\|}$$

Basic Algorithm

- Pick random initialization vectors
- Run power iterations
- Go with the winner, deflate and repeat

Algorithmic Improvements to Basic Tensor Method

$$\boxed{u \mapsto \frac{T(I, u, u)}{\|T(I, u, u)\|}}$$

Initialization Vectors

- Impacts convergence: different from matrix eigen-analysis.
- Random vectors: weak correlation with eigenvectors, noise sensitive.
- Whitened neighborhood vectors: Strong correlation with eigenvectors. More robust $\Rightarrow$ Better sample complexity.

Adaptive Deflation

- Better perturbation bounds for spectral estimation of tensors.

Algorithmic Improvements

Approaches(Anandkumar et al, COLT '13)

- Inverse moment method
- Preprocessing to whiten and symmetrize data
- Spectral approach: decompose tensor via batch power method
- Postprocessing: Recover Π from the spectrum by linear operations

Algorithmic Improvements

Approaches(Anandkumar et al, COLT '13)

- Inverse moment method
- Preprocessing to whiten and symmetrize data
- Spectral approach: decompose tensor via batch power method
- Postprocessing: Recover Π from the spectrum by linear operations

Parallelizable?	Speed?	Scalability?

Algorithmic Improvements

Approaches(Anandkumar et al, COLT '13)

- Inverse moment method
- Preprocessing to whiten and symmetrize data
- Spectral approach: decompose tensor via batch power method
- Postprocessing: Recover Π from the spectrum by linear operations

Parallelizable?	Speed?	Scalability?

Contribution Summary

- Randomized Low Rank Approximation for $n \times n$ matrix SVD
- Online tensor decomposition
- GPU Device to minimize data transfer overhead, thus fast updates
- Sparse Implementation scalable to millions of nodes
- Validation Metric: p-value test based "soft-pairing"

Stochastic (Implicit) Tensor Gradient Descent

$$\arg\min_{\mathbf{v}} \left\{ \left\| \theta \sum_{i \in [k]} \otimes^3 v_i - \sum_{t \in X} T^t \right\|_F^2 \right\},$$

where v_i are the unknown tensor eigenvectors, $T^t = g_A^t \otimes g_B^t \otimes g_C^t$ such that $g_A^t = W^\top G_{\{x,A\}}, \ldots$

Stochastic (Implicit) Tensor Gradient Descent

$$\arg\min_{\mathbf{v}} \left\{ \left\| \theta \sum_{i\in[k]} \otimes^3 v_i - \sum_{t\in X} T^t \right\|_F^2 \right\},$$

where v_i are the unknown tensor eigenvectors, $T^t = g_A^t \otimes g_B^t \otimes g_C^t$ such that $g_A^t = W^\top G_{\{x,A\}}, \ldots$

Expand the objective: $\theta \left\| \sum_{i\in[k]} \otimes^3 v_i \right\|_F^2 - \left\langle \sum_{i\in[k]} \otimes^3 v_i, T^t \right\rangle$

Orthogonality cost vs Correlation Reward

Stochastic (Implicit) Tensor Gradient Descent

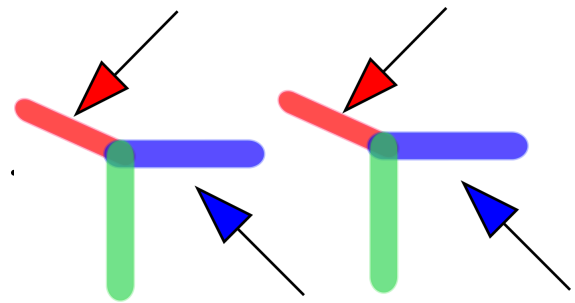
$$\arg\min_{\mathbf{v}} \left\{ \left\| \theta \sum_{i\in[k]} \otimes^3 v_i - \sum_{t\in X} T^t \right\|_F^2 \right\},$$

where v_i are the unknown tensor eigenvectors, $T^t = g_A^t \otimes g_B^t \otimes g_C^t$ such that $g_A^t = W^\top G_{\{x,A\}}, \ldots$

Expand the objective: $\theta \left\| \sum_{i\in[k]} \otimes^3 v_i \right\|_F^2 - \left\langle \sum_{i\in[k]} \otimes^3 v_i, T^t \right\rangle$

Orthogonality cost vs Correlation Reward

$$v_i^{t+1} \leftarrow v_i^t - 3\theta\beta^t \sum_{j=1}^{k} \left[\left\langle v_j^t, v_i^t \right\rangle^2 v_j^t \right] + \beta^t \left\langle v_i^t, g_A^t \right\rangle \left\langle v_i^t, g_B^t \right\rangle g_C^t + \ldots$$



Orthogonality cost vs Correlation Reward

Never form the tensor explicitly; multilinear operation on implicit tensor.

Computational Complexity ($k \ll n$)

- $n = \#$ of nodes
- $k = \#$ of communities
- $N = \#$ of iterations
- $m = \#$ of sampled node pairs (variational)

Module	Pre	STGD	Post	Var
Space	$O(nk)$	$O(k^2)$	$O(nk)$	$O(nk)$
Time	$O(n + k^3)$	$O(Nk)$	$O(n)$	$O(mkN)$

Variational method: $O(m \times k)$ **for each iteration**

$$O(n \times k) < O(m \times k) < O(n^2 \times k)$$

Our approach: $O(n + k^3)$

Computational Complexity ($k \ll n$)

- $n = \#$ of nodes
- $k = \#$ of communities
- $N = \#$ of iterations
- $m = \#$ of sampled node pairs (variational)

Module	Pre	STGD	Post	Var
Space	$O(nk)$	$O(k^2)$	$O(nk)$	$O(nk)$
Time	$O(n + k^3)$	$O(Nk)$	$O(n)$	$O(mkN)$

Variational method: $O(m \times k)$ **for each iteration**

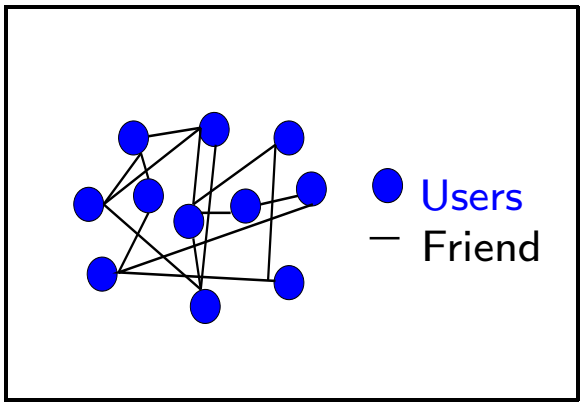
$$O(n \times k) < O(m \times k) < O(n^2 \times k)$$

Our approach: $O(n + k^3)$

In practice STGD is extremely fast and is not the bottleneck

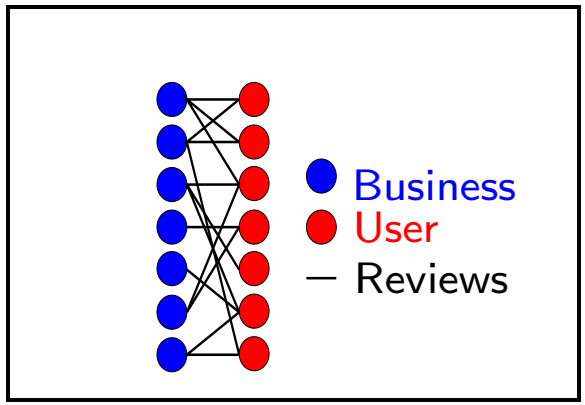
Outline

Summary of Results



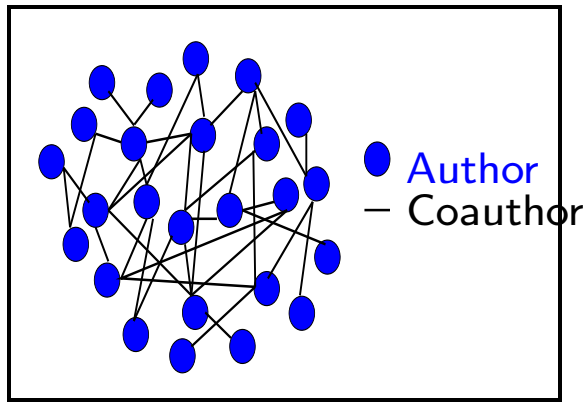
Facebook

$n \sim 20k$



Yelp

$n \sim 40k$



DBLP

$n \sim 1$ million

Error ($\mathcal{E}$) and Recovery ratio ($\mathcal{R}$)

Dataset	$\hat{k}$	Method	Running Time	$\mathcal{E}$	$\mathcal{R}$
Facebook(k=360)	500	ours	468	0.0175	100%
Facebook(k=360)	500	variational	86,808	0.0308	100%
Yelp(k=159)	100	ours	287	0.046	86%
Yelp(k=159)	100	variational	N.A.		
DBLP(k=6000)	100	ours	5407	0.105	95%

Summary of Results - Yelp Dataset

Lowest error business categories & largest weight businesses

Rank	Category	Business	Stars	Review Counts
1	Latin American	Salvadoreno Restaurant	4.0	36
2	Gluten Free	P.F. Chang's China Bistro	3.5	55
3	Hobby Shops	Make Meaning	4.5	14
4	Mass Media	KJZZ 91.5FM	4.0	13
5	Yoga	Sutra Midtown	4.5	31

Summary of Results - Yelp Dataset

Lowest error business categories & largest weight businesses

Rank	Category	Business	Stars	Review Counts
1	Latin American	Salvadoreno Restaurant	4.0	36
2	Gluten Free	P.F. Chang's China Bistro	3.5	55
3	Hobby Shops	Make Meaning	4.5	14
4	Mass Media	KJZZ 91.5FM	4.0	13
5	Yoga	Sutra Midtown	4.5	31

Bridgeness: Distance from vector $[1/\hat{k}, \ldots, 1/\hat{k}]^\top$

Top-5 bridging nodes (businesses)

Business	Categories
Four Peaks Brewing Co	Restaurants, Bars, American, Nightlife, Food, Pubs, Tempe
Pizzeria Bianco	Restaurants, Pizza, Phoenix
FEZ	Restaurants, Bars, American, Nightlife, Mediterranean, Lounges, Phoeni
Matt's Big Breakfast	Restaurants, Phoenix, Breakfast& Brunch
Cornish Pasty Company	Restaurants, Bars, Nightlife, Pubs, Tempe

Outline

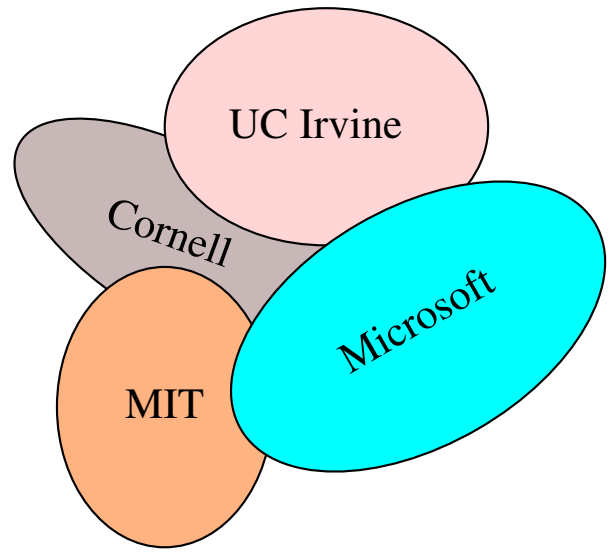
Conclusion

Mixed Membership Models

- Can model overlapping communities
- Efficient to learn from low order moments: edge counts and 3-star counts.

Tensor Spectral Method

- Whitened 3-star count tensor is an orthogonal symmetric tensor
- Efficient decomposition through power method
- Perturbation analysis: tight for stochastic block model
- Zero-error support recovery guarantees



Learning Overcomplete Representations

Tensor Approach to Learning Latent Variable Models

- Exploit conditional independence relations to obtain tensor form
- Low rank tensor when latent dim. $\ll$ observed dim.
- Applicable in community and document modeling

Learning Overcomplete Representations

Tensor Approach to Learning Latent Variable Models

- Exploit conditional independence relations to obtain tensor form
- Low rank tensor when latent dim. $\ll$ observed dim.
- Applicable in community and document modeling

Overcomplete Latent Representations

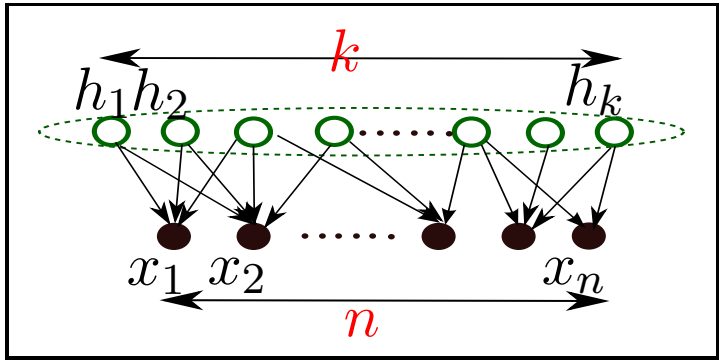
- Latent dimensionality $\gg$ observed dimensionality
- Flexible modeling, robust to noise
- Applicable in speech and image modeling

Novel Approaches for Learning Overcomplete Models

- Orthogonal tensor decomposition no longer applicable.
- Learning ill-posed for general models
- Solution: constraints through sparsity and/or incoherence

Two Approaches for Learning Overcomplete Models

Sparse bipartite graph Y.



Sparse Coding

- $h_1, \ldots, h_k$: Dictionary atoms
- $X = [x_1, \ldots, x_n]$: Observations
- Bipartite graph Y: Sparse mixing $X = HY$.
- Incoherent dictionary
- Clustering and alt. min.

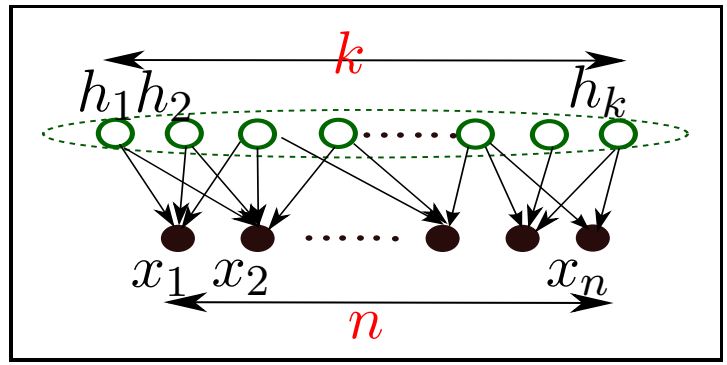
---

[1] A. Agarwal, A., P. Netrapalli. "Exact Recovery of Sparsely Used Overcomplete Dictionaries," Preprint, Sept. 2013.

[2] A., D. Hsu, M. Janzamin, and S. M. Kakade. When are Overcomplete Representations Identifiable? Uniqueness of Tensor Decompositions Under Expansion Constraints, NIPS, Dec. 2013.

Two Approaches for Learning Overcomplete Models

Sparse bipartite graph Y.



Sparse Coding

- $h_1, \ldots, h_k$: Dictionary atoms
- $X = [x_1, \ldots, x_n]$: Observations
- Bipartite graph Y: Sparse mixing $X = HY$.
- Incoherent dictionary
- Clustering and alt. min.

Sparse Topic Models

- $h_1, \ldots, h_k$: Topics
- $X = [x_1, \ldots, x_n]$: word. Three words/view X_1, X_2, X_3.
- Y: Topic-word matrix. $\mathbb{E}[X_1 \otimes X_2 \otimes X_3] = \mathbb{E}[H \otimes H \otimes H](Y, Y, Y$
- Multi-view and Persistent topics
- Tensor decomposition

---

[1] A. Agarwal, A., P. Netrapalli. "Exact Recovery of Sparsely Used Overcomplete Dictionaries," Preprint, Sept. 2013.

[2] A., D. Hsu, M. Janzamin, and S. M. Kakade. When are Overcomplete Representations Identifiable? Uniqueness of Tensor Decompositions Under Expansion Constraints, NIPS, Dec. 2013.