The University of Texas at Austin
Department of Computer Science
College of Natural Sciences

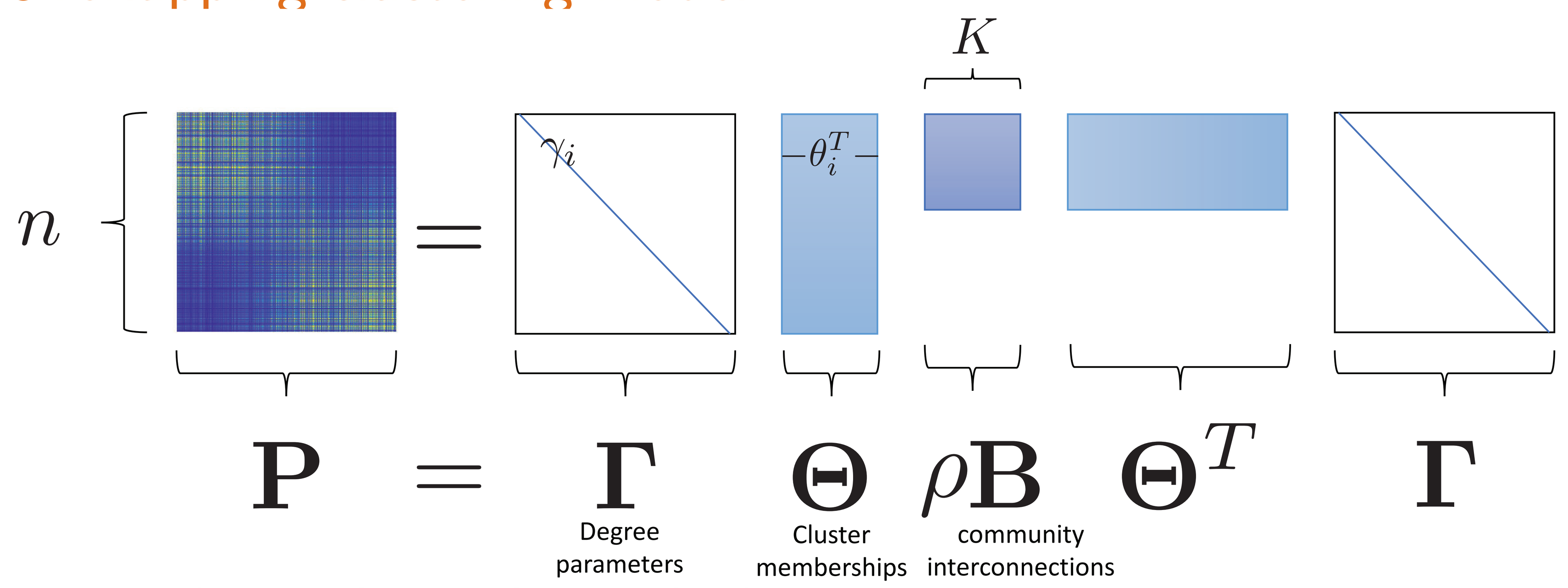
Overlapping Clustering Models, and One (class) SVM to Bind Them All

Xueyu Mao, Purnamrita Sarkar, Deepayan Chakrabarti

The University of Texas at Austin
Department of Statistics and Data Sciences
College of Natural Sciences

The University of Texas at Austin
McCombs School of Business

Overlapping clustering model



- Largest element of $\mathbf{B}$ is 1 for identifiability
- For network models, data is adjacency matrix: $\mathbf{A}_{ij} \sim \text{Bernoulli}(\mathbf{P}_{ij})$
- Refer as "DCMMSB-type models"
  - Special cases:
    - $\|\boldsymbol{\theta}_i\|_1 = 1$, Degree-corrected Mixed Membership Stochastic Blockmodel (DCMMSB) (Airoldi et al., 2008; Jin et al., 2017)
    - $\|\boldsymbol{\theta}_i\|_2 = 1$, Overlapping Continuous Community Assignment Model (OCCAM) (Zhang et al. 2014)
    - $\boldsymbol{\theta}_i$ binary, may have multiple non-zero entries, $\mathbf{\Gamma}$ identity, Stochastic Blockmodel with Overlaps (SBMO) (Kaufmann et al. 2016)
- For topic models (Blei et al. 2003), data is word count matrix for different documents coming from multinomial sample of the population
  - Word co-occurrence probability matrix matches the form of $\mathbf{P}$ in network models after normalizing the rows of word-topic matrix

Related work

Most algorithms (Mao et al., 2017a, 2017b; Jin et al., 2017; Panov et al., 2017; Rubin-Delanchy et al., 2017) use a two step method:

- Find the pure nodes by finding corners of a **simplex**
- Estimate model parameters via regression

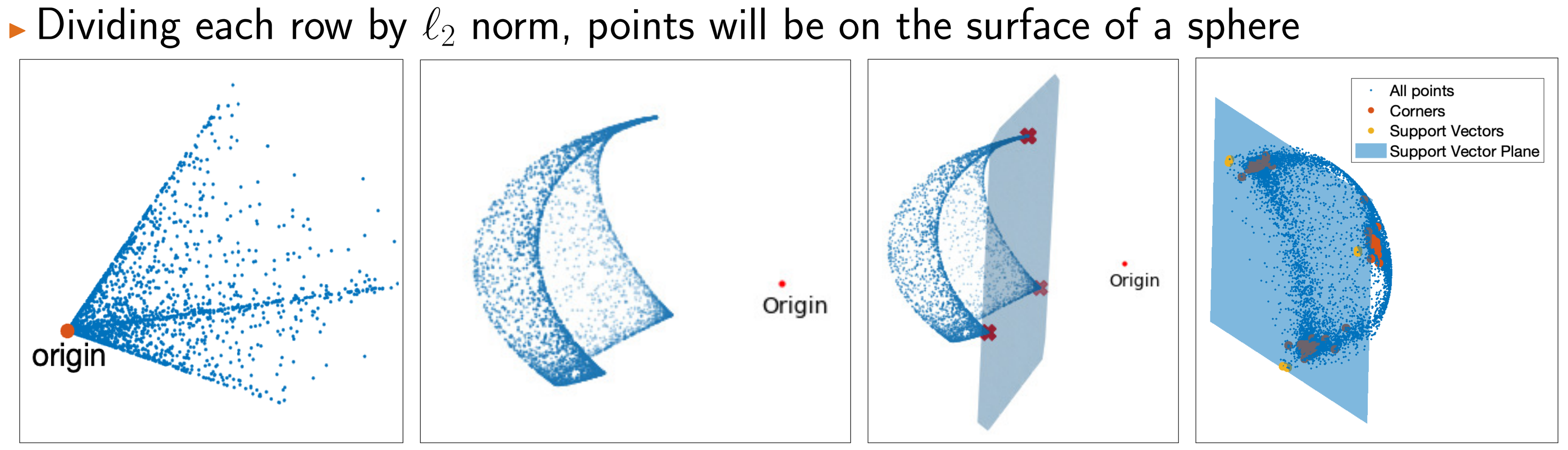
For example, (Mao et al., 2017b) shows that for MMSB,

- Rows $\mathbf{v}_i$ of $\mathbf{V}$ (the top-K eigenvectors of $\mathbf{P}$) lie on a simplex with corners as pure nodes (set $\mathcal{I}$)
- One can estimate community memberships by regressing $\mathbf{v}_i$ on $\mathbf{V}(\mathcal{I},:)$

With degree parameters, rows of $\mathbf{V}$ are not in a simplex.

Eliminating the effects of degree parameters

- Rows of $\mathbf{V}$ fall on a **cone**
- Dividing each row by ℓ_2 norm, points will be on the surface of a sphere



- A hyperplane through the corners separates all the points from the origin
  - Can find the hyperplane with a One-class SVM (Schölkopf et al, 2001), support vectors are the corners

Ideal cone problem

- Given a matrix $\mathbf{Z}$ that is known to be of the form $\mathbf{Z} = \mathbf{M}\mathbf{Y}_P \in \mathbb{R}^{n\times m}$, where
  - $\mathbf{M} \in \mathbb{R}_{\geq 0}^{n\times K}$, no row of $\mathbf{M}$ is 0
  - $\mathbf{Y}_P \in \mathbb{R}^{K\times m}$ corresponds to K (unknown) rows of $\mathbf{Z}$, each scaled to unit ℓ_2 norm
- Infer $\mathbf{M}$

Solving the ideal cone problem with $\mathbf{Z} = \mathbf{V}$ gives us the pure nodes, from which community memberships can be inferred.

One-class SVM solves ideal cone problem

Primal:

$$\max \quad b \qquad s.t. \quad \|\mathbf{w}\| \leq 1,\ \mathbf{w}^T\mathbf{y}_i \geq b,\ i \in \mathcal{I}$$

Dual:

$$\min \quad \frac{1}{2}\sum_{i,j}\beta_i\beta_j\mathbf{y}_i^T\mathbf{y}_j \qquad s.t. \quad \sum_i \beta_i = 1,\ \beta_i \geq 0,\ i \in \mathcal{I}$$

- where $\mathbf{y}_i = \mathbf{z}_i/\|\mathbf{z}_i\|$
- One-class SVM works if the following condition is satisfied:

Condition

The matrix $\mathbf{Y}_P$ satisfies $(\mathbf{Y}_P\mathbf{Y}_P^T)^{-1}\mathbf{1} > \mathbf{0}$.

- Intuition: the SVM support vector plane touches all pure nodes
- **Condition always holds for DCMMSB-type models**

Empirical Cone Problem

- Given the empirical matrix $\hat{\mathbf{Y}}$ with rows $\hat{\mathbf{z}}_i^T/\|\hat{\mathbf{z}}_i\|$
  - with $\mathbf{Z} = \mathbf{M}\mathbf{Y}_P \in \mathbb{R}^{n\times m}$, and
  - $\max_i \|\mathbf{e}_i^T(\hat{\mathbf{Y}} - \mathbf{Y})\| \leq \epsilon$.
- Infer $\hat{\mathbf{M}}$

One-class SVM still works if the following condition is satisfied:

Condition

The matrix $\mathbf{Y}_P$ satisfies $(\mathbf{Y}_P\mathbf{Y}_P^T)^{-1}\mathbf{1} \geq \eta\mathbf{1}$ for some constant $\eta > 0$.

- η can be bounded for DCMMSB-type models and topic models

Algorithm

SVM-cone

- Normalize rows of $\hat{\mathbf{Z}}$ by ℓ_2 norm
- Run one-class SVM to get supporting hyperplane
- Cluster all points close to this hyperplane
- Pick one point from each cluster to get near-corner set C
- $\hat{\mathbf{M}} = \hat{\mathbf{Z}}\hat{\mathbf{Y}}_C^T(\hat{\mathbf{Y}}_C\hat{\mathbf{Y}}_C^T)^{-1}$

Proof Sketch for Empirical Cone Problem

- Step 1: Show that SVM solution of empirical cone is nearly ideal, i.e., $(\hat{\mathbf{w}}, \hat{b}) \approx (\mathbf{w}, b)$
- Step 2: Show that true corners of the cone are close to supporting hyper plane
- Step 3: Also, all points close to support vectors are nearly corners
- Step 4: Clustering the points that are close to support vectors yields exactly one cluster for each corner

Per-node Consistency

- If $\|\mathbf{y}_i - \hat{\mathbf{y}}_i\| \leq \epsilon \to 0$ for all i, using SVM-cone will get $\mathbf{M}$ consistently with ground truth ($\|\mathbf{m}_i - \hat{\mathbf{m}}_i\| \to 0$)
- Holds with high probability for both DCMMSB-type models and topic models

Per-node consistency guarantee for DCMMSB (informal)

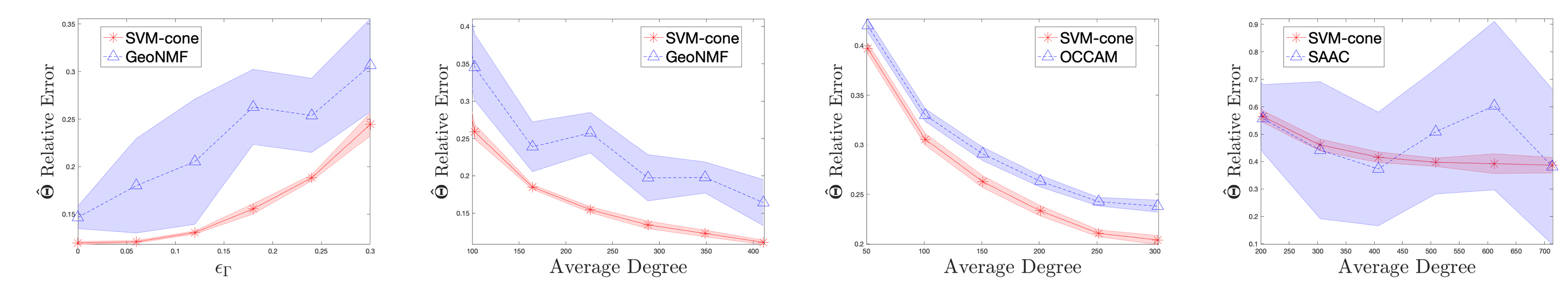
If $\boldsymbol{\theta}_i \sim \text{Dirichlet}(\boldsymbol{\alpha})$, $\alpha_0 = \boldsymbol{\alpha}^T\mathbf{1}$, under some conditions on $\boldsymbol{\alpha}$ and $\mathbf{\Gamma}$, w.h.p.,

$$\|\mathbf{e}_i^T(\mathbf{\Theta} - \hat{\mathbf{\Theta}}\mathbf{\Pi})\| = \tilde{O}\left(\frac{\gamma_{\max}K^{2.5}\min\{K^2, (\kappa(\mathbf{P}))^2\}\nu^2(1+\alpha_0)^2}{\gamma_{\min}^5\eta\lambda^*(\mathbf{B})\sqrt{\rho n}}\right)$$

- $\lambda^*(\mathbf{B})$: smallest singular value of $\mathbf{B}$, controls the separation between clusters
- ν: controls how balanced entries of $\boldsymbol{\alpha}$ are
- **Similar per-node consistency guarantees for all DCMMSB-type models, and per-word guarantees for topic models.**

Simulation Experiments

- Comparing with GeoNMF (Mao et al. 2017), OCCAM (Zhang et al. 2014) and SAAC (Kaufmann et al. 2016)



(a) (b) (c) (d)

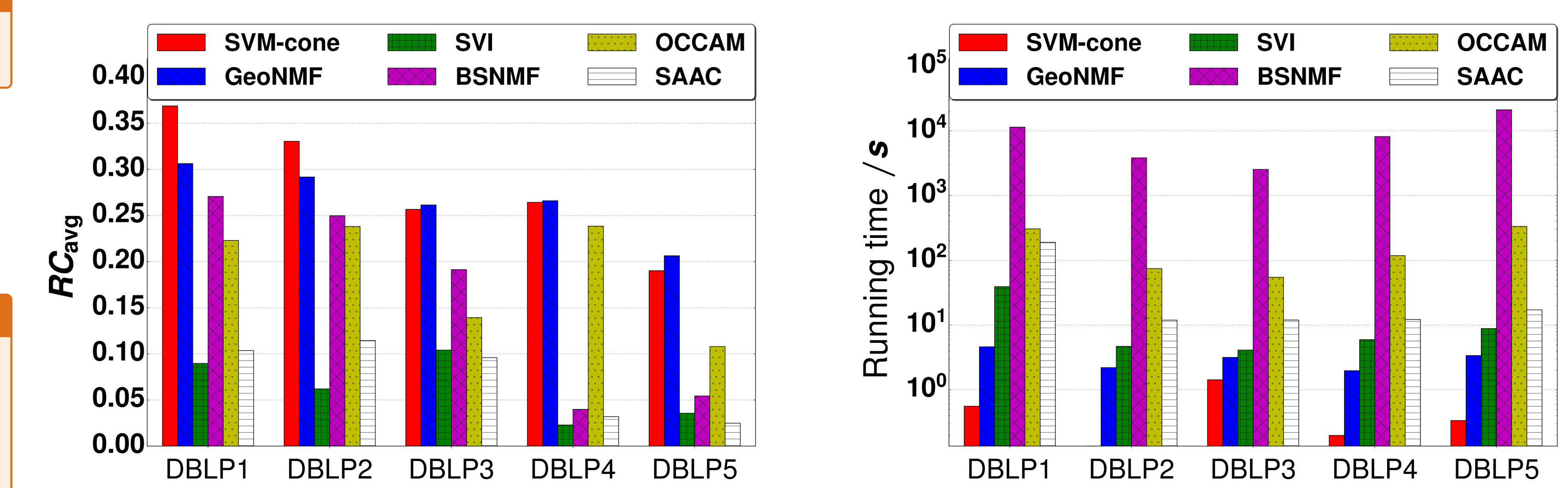
(a) Varying degree heterogeneity on DCMMSB. (b) Varying sparsity on DCMMSB. (c) Varying sparsity on OCCAM. (d) Varying sparsity on SBMO.

Real-world Network Results

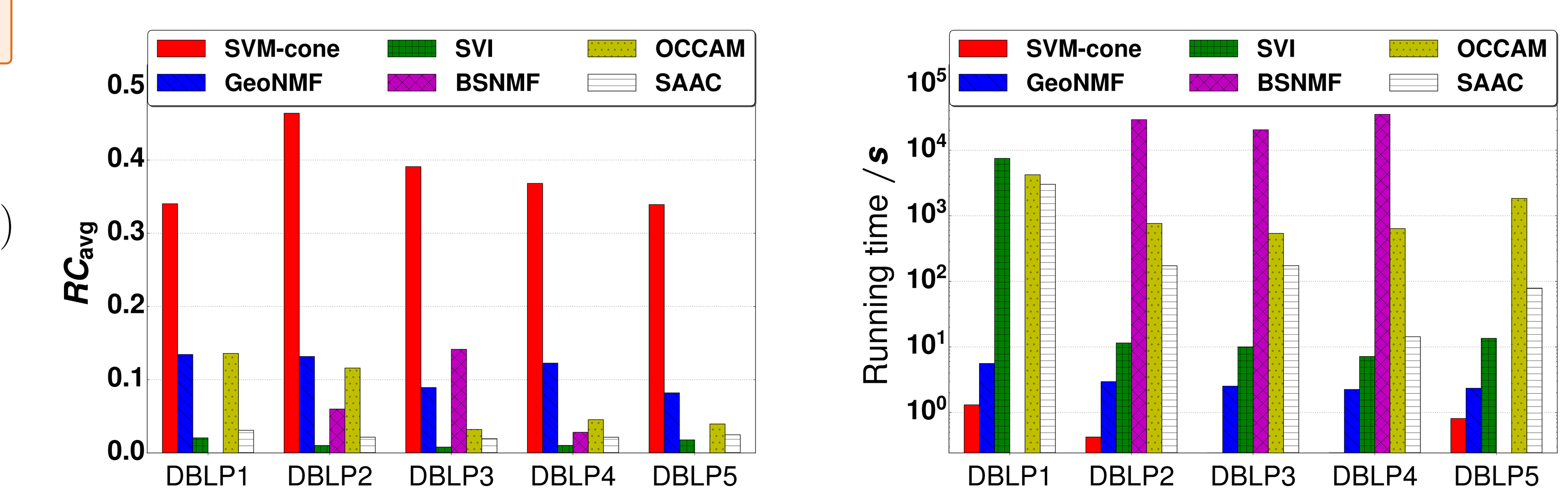
- Evaluation metric: Averaged Spearman rank correlation coefficients (RC) between $\mathbf{\Theta}(:,a)$, $a \in [K]$ and $\hat{\mathbf{\Theta}}(:,\sigma(a))$, where σ is a permutation of $[K]$.

$$\text{RC}_{\text{avg}}(\hat{\mathbf{\Theta}}, \mathbf{\Theta}) = \frac{1}{K}\max_{\sigma}\sum_{i=1}^{K}\text{RC}(\hat{\mathbf{\Theta}}(:,i), \mathbf{\Theta}(:,\sigma(i))) \in [-1, 1].$$

- Co-authorship networks (assortative)



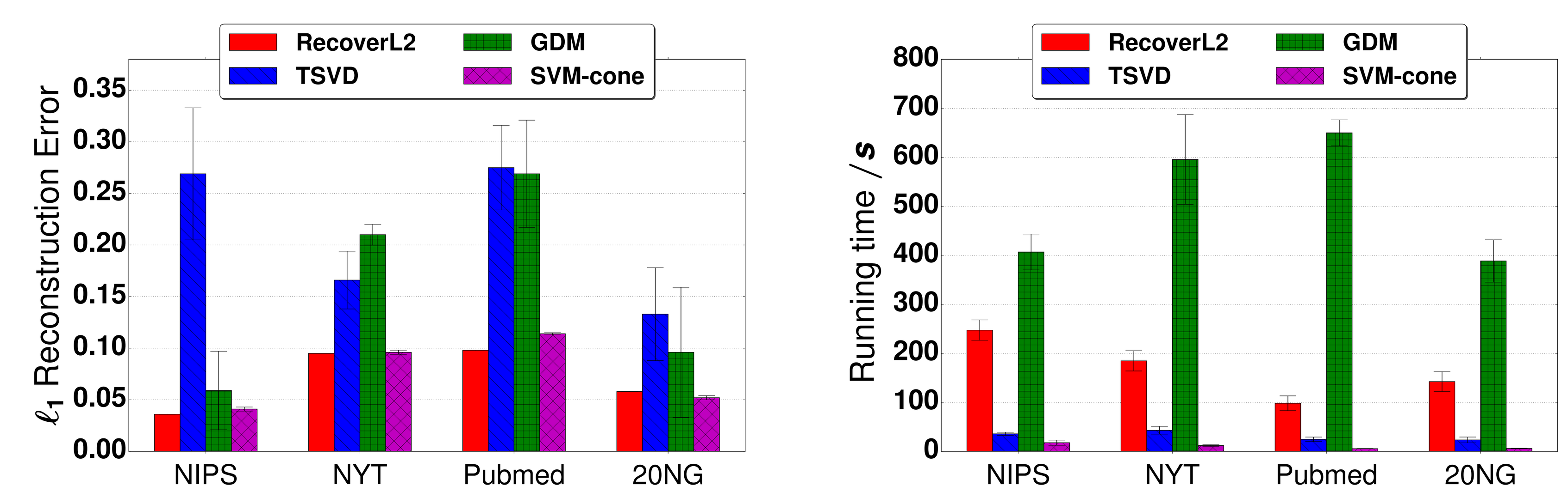
- Author-paper bipartite networks (dissortative)



(a) Rank correlation. (b) Running time (log scale) .

Topic Models Results

- ℓ_1 reconstruction error and running time (log scale) for semi-synthetic data with number of documents set to 60,000



{xmao@cs, purna.sarkar@austin, deepay@}.utexas.edu