Algorithmic Advances for the Adjacency Spectral Embedding

Gonzalo Mateos

Dept. of ECE and Goergen Institute for Data Science
University of Rochester
gmateosb@ece.rochester.edu
http://www.hajim.rochester.edu/ece/sites/gmateos

Co-authors: M. Fiori, B. Marenco, F. Larroca, and P. Bermolen

Acknowledgment: NSF Awards CCF-1750428, CCF-1934962 and ECCS-1809356

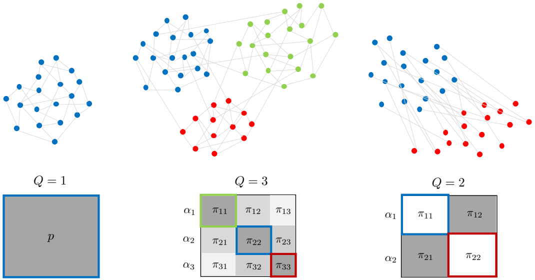
EUSIPCO 2022, Belgrade, Serbia
August 30, 2022

Classical random graph models

- Simplest model is the Erdös-Renyi (ER) random graph model $G_{N_v,p}$
  - $\Rightarrow$ Undirected graph N_v vertices, edges present w.p. p, independently

- Simplest model is the Erdös-Renyi (ER) random graph model $G_{N_v,p}$
  - $\Rightarrow$ Undirected graph N_v vertices, edges present w.p. p, independently
- Stochastic block models (SBMs) explicitly parameterize the notion of communities $\mathcal{C}_1, \ldots, \mathcal{C}_Q$
  - $\Rightarrow$ Nodes assigned to $\mathcal{C}_q$ w.p. α_q, connection rates π_{qr} of vertices between/within groups

- Simplest model is the Erdös-Renyi (ER) random graph model $G_{N_v,p}$
  - $\Rightarrow$ Undirected graph N_v vertices, edges present w.p. p, independently
- Stochastic block models (SBMs) explicitly parameterize the notion of communities $\mathcal{C}_1, \ldots, \mathcal{C}_Q$
  - $\Rightarrow$ Nodes assigned to $\mathcal{C}_q$ w.p. α_q, connection rates π_{qr} of vertices between/within groups



- Mixtures of Erdős-Rényi models can be surprisingly flexible

Random dot product graphs

- Consider a latent space $\mathcal{X}_d \subset \mathbb{R}^d$ such that for all

$$\mathbf{x}, \mathbf{y} \in \mathcal{X}_d \quad \Rightarrow \quad \mathbf{x}^\top \mathbf{y} \in [0, 1]$$

$\Rightarrow$ Inner-product distribution $F : \mathcal{X}_d \mapsto [0, 1]$

- Random dot product graphs (RDPGs) are defined as follows:

$$\mathbf{x}_1, \ldots, \mathbf{x}_{N_v} \overset{\text{i.i.d.}}{\sim} F,$$
$$A_{ij} \,\big|\, \mathbf{x}_i, \mathbf{x}_j \sim \text{Bernoulli}(\mathbf{x}_i^\top \mathbf{x}_j)$$

for $1 \leq i, j \leq N_v$, where $A_{ij} = A_{ji}$ and $A_{ii} \equiv 0$

- A particularly tractable latent position random graph model

$\Rightarrow$ Vertex positions $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_{N_v}]^\top \in \mathbb{R}^{N_v \times d}$

S. J. Young and E. R. Scheinerman, "Random dot product graph models for social networks," *WAW*, 2007

- RDPGs ecompass the previous classic models for network graphs

  Ex: Recover Erdös-Renyi $G_{N_v,p}$ graphs with $d = 1$ and $\mathcal{X}_d = \{\sqrt{p}\}$

  Ex: Recover SBM random graphs by constructing F with pmf

$$\mathrm{P}\left(\mathbf{X} = \mathbf{x}_q\right) = \alpha_q, \quad q = 1, \ldots, Q$$

  after selecting d and $\mathbf{x}_1, \ldots, \mathbf{x}_Q$ such that $\pi_{qr} = \mathbf{x}_q^\top \mathbf{x}_r$

- Approximation results for SBMs justify the expressiveness of RDPGs

Estimation of latent positions

- **Q:** Given G from an RDPG, find the 'best' $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_{N_v}]^\top$?
- MLE is well motivated but it is intractable for large N_v

$$\hat{\mathbf{X}}_{ML} = \underset{\mathbf{X}}{\text{argmax}} \prod_{i<j} (\mathbf{x}_i^\top \mathbf{x}_j)^{A_{ij}} (1 - \mathbf{x}_i^\top \mathbf{x}_j)^{1-A_{ij}}$$

- **Q:** Given G from an RDPG, find the 'best' $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_{N_v}]^\top$?
- MLE is well motivated but it is intractable for large N_v

$$\hat{\mathbf{X}}_{ML} = \underset{\mathbf{X}}{\text{argmax}} \prod_{i<j} (\mathbf{x}_i^\top \mathbf{x}_j)^{A_{ij}} (1 - \mathbf{x}_i^\top \mathbf{x}_j)^{1-A_{ij}}$$

- Instead, let $P_{ij} = \mathrm{P}\left((i,j) \in \mathcal{E}\right)$ and define $\mathbf{P} = [P_{ij}] \in [0,1]^{N_v \times N_v}$
  - $\Rightarrow$ The RDPG model specifies that $\mathbf{P} = \mathbf{X}\mathbf{X}^\top$
  - $\Rightarrow$ **Key:** Observed $\mathbf{A}$ is a noisy realization of $\mathbf{P}$ $(\mathbb{E}[\mathbf{A}] = \mathbf{P})$
- Suggests a LS regression approach to find $\mathbf{X}$ s.t. $\mathbf{X}\mathbf{X}^\top \approx \mathbf{A}$

$$\hat{\mathbf{X}}_{LS} = \underset{\mathbf{X}}{\text{argmin}} \|\mathbf{X}\mathbf{X}^\top - \mathbf{A}\|_F^2$$

A. Athreya et al, "Statistical inference on random dot product graphs: A survey," *JMLR*, 2018

Adjacency spectral embedding

- Since $\mathbf{A}$ is real and symmetric, can decompose it as $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$
  - $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_{N_v}]$ is the orthogonal matrix of eigenvectors
  - $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_{N_v})$, with eigvenvalues $\lambda_1 \geq \ldots \geq \lambda_{N_v}$

Adjacency spectral embedding

- Since $\mathbf{A}$ is real and symmetric, can decompose it as $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$
  - $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_{N_v}]$ is the orthogonal matrix of eigenvectors
  - $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_{N_v})$, with eigvenvalues $\lambda_1 \geq \ldots \geq \lambda_{N_v}$
- Define $\hat{\mathbf{\Lambda}} = \text{diag}(\lambda_1^+, \ldots, \lambda_d^+)$ and $\hat{\mathbf{U}} = [\mathbf{u}_1, \ldots, \mathbf{u}_d]$ $(\lambda^+ := \max(0, \lambda))$
- Best rank-d, positive semi-definite (PSD) approximation of $\mathbf{A}$ is $\hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}^\top$

$\Rightarrow$ Ajacency spectral embedding (ASE) is $\hat{\mathbf{X}}_{LS} = \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}^{1/2}$ since

$$\mathbf{A} \approx \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}^\top = \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}^{1/2}\hat{\mathbf{\Lambda}}^{1/2}\hat{\mathbf{U}}^\top = \hat{\mathbf{X}}_{LS}\hat{\mathbf{X}}_{LS}^\top$$

- Since $\mathbf{A}$ is real and symmetric, can decompose it as $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$
  - $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_{N_v}]$ is the orthogonal matrix of eigenvectors
  - $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_{N_v})$, with eigvenvalues $\lambda_1 \geq \ldots \geq \lambda_{N_v}$
- Define $\hat{\mathbf{\Lambda}} = \text{diag}(\lambda_1^+, \ldots, \lambda_d^+)$ and $\hat{\mathbf{U}} = [\mathbf{u}_1, \ldots, \mathbf{u}_d]$ $(\lambda^+ := \max(0, \lambda))$
- Best rank-d, positive semi-definite (PSD) approximation of $\mathbf{A}$ is $\hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}^\top$

  $\Rightarrow$ Ajacency spectral embedding (ASE) is $\hat{\mathbf{X}}_{LS} = \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}^{1/2}$ since

$$\mathbf{A} \approx \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}^\top = \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}^{1/2}\hat{\mathbf{\Lambda}}^{1/2}\hat{\mathbf{U}}^\top = \hat{\mathbf{X}}_{LS}\hat{\mathbf{X}}_{LS}^\top$$

- **Q:** Is the solution unique? Nope, inner-products are rotation invariant

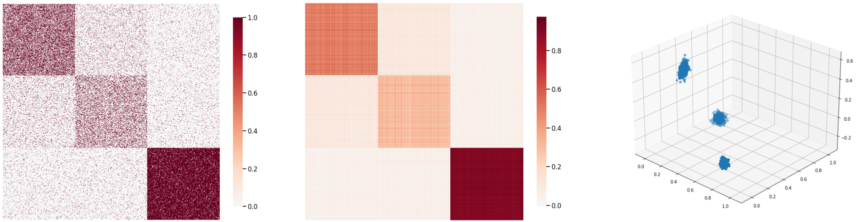
$$\mathbf{P} = \mathbf{X}\mathbf{W}(\mathbf{X}\mathbf{W})^\top = \mathbf{X}\mathbf{X}^\top, \quad \mathbf{W}\mathbf{W}^\top = \mathbf{I}_d$$

  $\Rightarrow$ RDPG embedding problem is identifiable modulo rotations

Embedding an SBM graph

- Ex: SBM with $N_v = 1500$, $Q = 3$ and mixing parameters

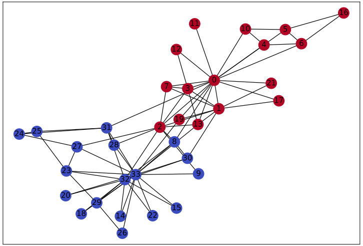
$$\boldsymbol{\alpha} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}, \quad \boldsymbol{\Pi} = \begin{bmatrix} 0.5 & 0.1 & 0.05 \\ 0.1 & 0.3 & 0.05 \\ 0.05 & 0.05 & 0.9 \end{bmatrix}$$

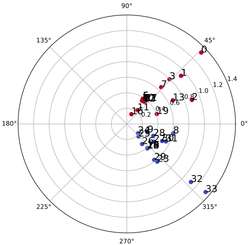


- Sample adjacency $\mathbf{A}$ (left), $\hat{\mathbf{X}}_{LS}\hat{\mathbf{X}}_{LS}^{\top}$ (center), rows of $\hat{\mathbf{X}}_{LS}$ (right)
- Use embeddings to bring to bear geometric methods of analysis

Interpretability of the embeddings

- Ex: Zachary's karate club graph with $N_v = 34$, $N_e = 78$ (left)





- Node embeddings (rows of $\hat{\mathbf{X}}_{LS}$) for $d = 2$ (right)
  - Club's administrator ($i = 0$) and instructor ($j = 33$) are orthogonal
- Interpretability of embeddings a valuable asset for RDPGs

  $\Rightarrow$ Vector magnitudes indicate how well connected nodes are

  $\Rightarrow$ Vector angles indicate positions in latent space

Limitations and open questions

- `Graspologic`: SoA package for RDPG estimation in Python (eigendecompostion in `SciPy`)
  Q: Scalability for large graphs? Streaming settings for dynamic graphs?
- ASE does not account for the all-zero diagonal of $\mathbf{A}$. Truly we want to solve

$$\hat{\mathbf{X}} \in \operatorname*{argmin}_{\mathbf{X} \in \mathbb{R}^{N \times d}} \|\mathbf{M} \circ (\mathbf{A} - \mathbf{X}\mathbf{X}^\top)\|_F^2$$

  $\Rightarrow \mathbf{M} := \mathbf{1}\mathbf{1}^\top - \mathbf{I}$ is a mask matrix, with zero-diagonal and ones everywhere else
- Iterative approach in [Scheinerman-Tucker'10] is complex and may not converge
  Q: First-order non-convex optimization methods? Missing data in $\mathbf{A}$?

Limitations and open questions

- `Graspologic`: SoA package for RDPG estimation in Python (eigendecompostion in `SciPy`)

  Q: Scalability for large graphs? Streaming settings for dynamic graphs?

- ASE does not account for the all-zero diagonal of $\mathbf{A}$. Truly we want to solve

$$\hat{\mathbf{X}} \in \underset{\mathbf{X} \in \mathbb{R}^{N \times d}}{\operatorname{argmin}} \|\mathbf{M} \circ (\mathbf{A} - \mathbf{X}\mathbf{X}^\top)\|_F^2$$

  $\Rightarrow$ $\mathbf{M} := \mathbf{1}\mathbf{1}^\top - \mathbf{I}$ is a mask matrix, with zero-diagonal and ones everywhere else

- Iterative approach in [Scheinerman-Tucker'10] is complex and may not converge

  Q: First-order non-convex optimization methods? Missing data in $\mathbf{A}$?

- **Goal:** an RDPG inference approach that offers a better representation at a lower computational cost, in more general settings

- Let $f : \mathbb{R}^{N_v \times d} \mapsto \mathbb{R}$ be the objective function $f(\mathbf{X}) = \|\mathbf{M} \circ (\mathbf{A} - \mathbf{X}\mathbf{X}^\top)\|_F^2$

  $\Rightarrow$ Non-convex w.r.t. $\mathbf{X}$, convex w.r.t. $\mathbf{P} = \mathbf{X}\mathbf{X}^\top$

- Gradient descent (GD) method (a.k.a. *factorized GD* or *Procrustes flow*)

$$\mathbf{X}_{t+1} = \mathbf{X}_t - \alpha \nabla f(\mathbf{X}_t), \quad t = 0, 1, 2, \ldots$$

  $\Rightarrow$ Step size $\alpha > 0$ and $\nabla f(\mathbf{X}) = 4\left[\mathbf{M} \circ (\mathbf{X}\mathbf{X}^\top - \mathbf{A})\right]\mathbf{X}$, for symmetric $\mathbf{A}$ and $\mathbf{M}$

Gradient descent

- Let $f : \mathbb{R}^{N_v \times d} \mapsto \mathbb{R}$ be the objective function $f(\mathbf{X}) = \|\mathbf{M} \circ (\mathbf{A} - \mathbf{X}\mathbf{X}^\top)\|_F^2$

  $\Rightarrow$ Non-convex w.r.t. $\mathbf{X}$, convex w.r.t. $\mathbf{P} = \mathbf{X}\mathbf{X}^\top$

- Gradient descent (GD) method (a.k.a. *factorized GD* or *Procrustes flow*)

$$\mathbf{X}_{t+1} = \mathbf{X}_t - \alpha \nabla f(\mathbf{X}_t), \quad t = 0, 1, 2, \ldots$$

  $\Rightarrow$ Step size $\alpha > 0$ and $\nabla f(\mathbf{X}) = 4\left[\mathbf{M} \circ (\mathbf{X}\mathbf{X}^\top - \mathbf{A})\right]\mathbf{X}$, for symmetric $\mathbf{A}$ and $\mathbf{M}$

- Convergence: if $\mathbf{X}_0$ is close to the solution, the iteration converges with linear rate to $\hat{\mathbf{X}}$

Proposition. There exist $\delta > 0$ and $0 < \kappa < 1$ such that, if $\|\mathbf{X}_0 - \hat{\mathbf{X}}\|_F \leq \delta$, then

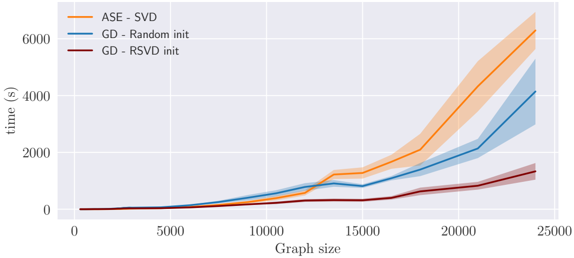
$$d(\mathbf{X}_t, \hat{\mathbf{X}}) \leq \kappa^t d(\mathbf{X}_0, \hat{\mathbf{X}}), \quad \text{for all } t > 0,$$

where $d(\mathbf{X}, \hat{\mathbf{X}}) := \min_{\mathbf{W} \in \mathcal{O}^{d \times d}} \|\mathbf{X}\mathbf{W} - \hat{\mathbf{X}}\|_F^2$ accounts for the rotational ambiguity.

Y. Chi *et al*, "Nonconvex optimization meets low-rank matrix factorization: An overview," *TSP*, 2019

Numerical tests

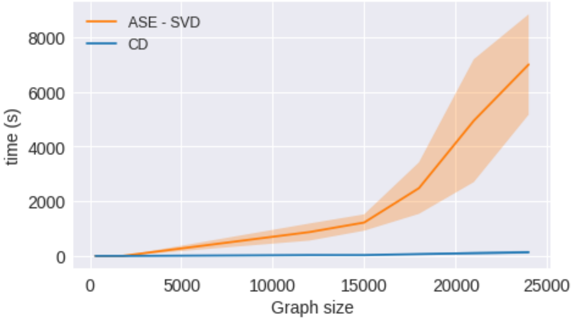
- Scalability to large graphs



Run-time comparison for SBMs with up to $N_v = 25000$ nodes

Numerical tests

- Scalability to large graphs



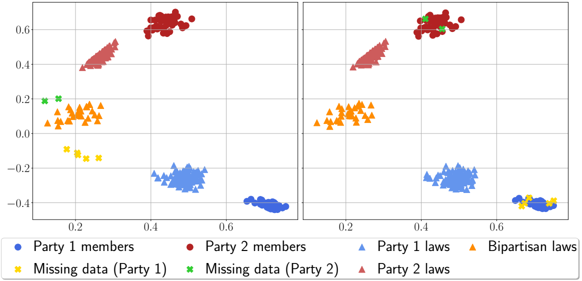
Even better with coordinate descent approaches

Numerical tests

- Scalability to large graphs
- Better quality representations

Since we discard residuals for the diagonal entries of $\mathbf{A}$

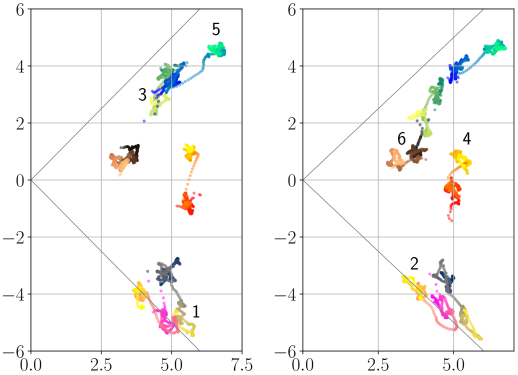
- Scalability to large graphs
- Better quality representations
- Can handle missing data



Bipartite graph: parties and laws

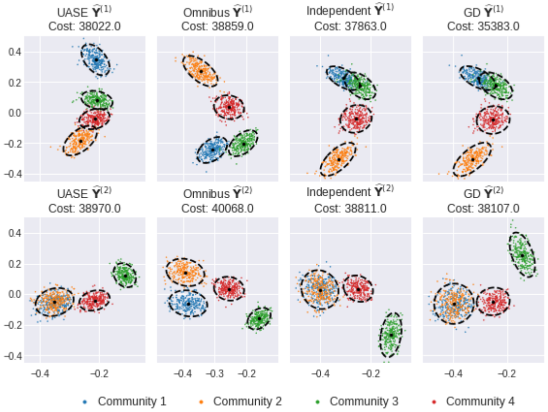
- Parties submit their laws
- Some bipartisan laws
- Some senators have 30% chance of being absent

- Scalability to large graphs
- Better quality representations
- Can handle missing data
- Tracking via warm restarts



- $N_v = 6$ Wi-Fi APs in a rural Uruguayan school
- Hourly measurements over 4 weeks (655 graphs)
- AP 4 was moved at $t \approx 310$

- Scalability to large graphs
- Better quality representations
- Can handle missing data
- Tracking via warm restarts



- Two SBMs with $Q = 4$ communities, $\mathbf{\Pi}$ changes
- Communities 1 and 2 merge in the second model
- Community 4 maintains probabilities w/other groups

Gallagher *et al*, "Spectral embedding for dynamic networks with stability guarantees," *NeurIPS*, 2021

Concluding remarks

- ASE to estimate latent nodal positions in RDPGs $\Rightarrow$ Non-convex matrix factorization problem
- Convergent, first-order gradient descent algorithm for refined formulation
  - $\Rightarrow$ Scalable and fast computation of nodal representations
  - $\Rightarrow$ Can handle missing data
  - $\Rightarrow$ Able to track representation for dynamic graphs
- Try it @ `https://github.com/git-artes/cpd_rdpg`
- Future work
  - $\Rightarrow$ Embedding dynamic network with varying number of nodes (Asilomar'22)
  - $\Rightarrow$ Statistical properties such as consistency, asymptotic normality, stability ($N_v \to \infty$)