MuDCoD

Multi-Subject Community Detection in Dynamic Gene Networks

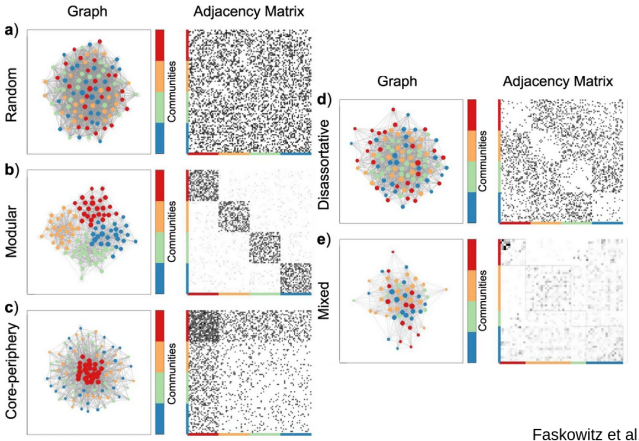
Ali Osman Berk Şapcı

Outline

- Community detection and spectral clustering
- Gene co-expression networks and scRNA-seq datasets
- Multi-subject dynamic networks and MuDCoD
- Evaluating MuDCoD

Detecting communities in a network

Subsets of nodes which are densely connected internally and sparsely connected externally.

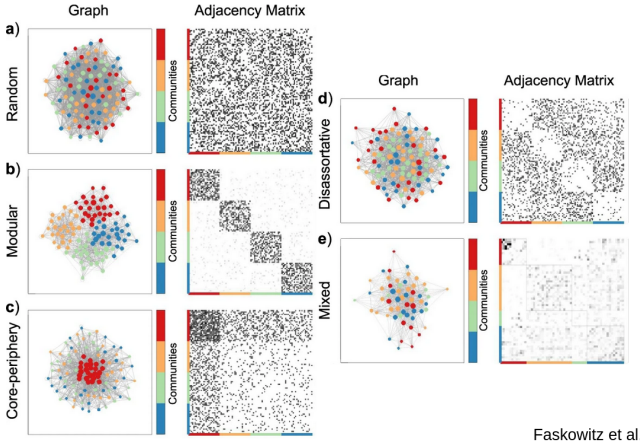


Faskowitz et al., 2018

Detecting communities in a network

Subsets of nodes which are densely connected internally and sparsely connected externally.

- Can be overlapping or non-overlapping.

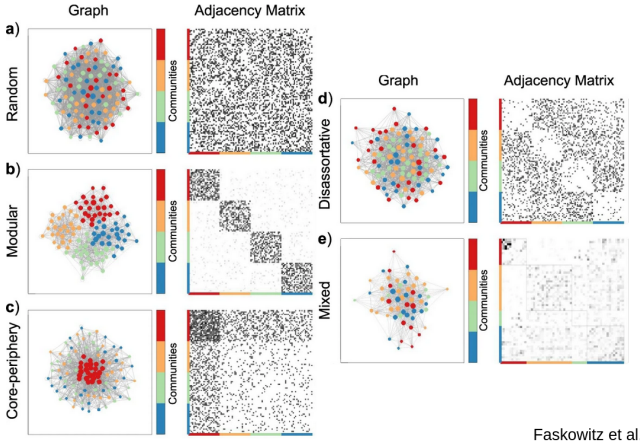


Faskowitz et al., 2018

Detecting communities in a network

Subsets of nodes which are densely connected internally and sparsely connected externally.

- Can be overlapping or non-overlapping.
- Some networks may not have any meaningful community structure.

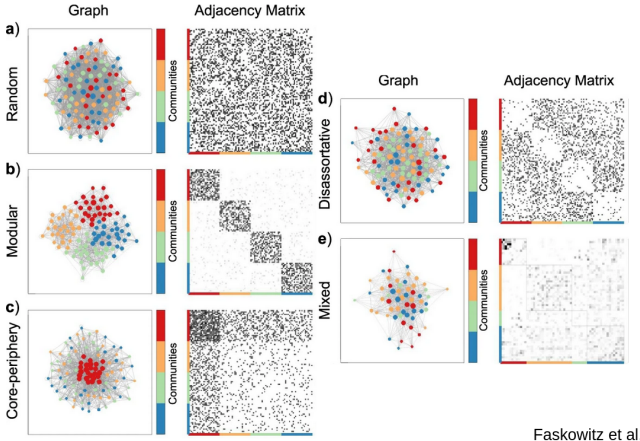


Faskowitz et al., 2018

Detecting communities in a network

Subsets of nodes which are densely connected internally and sparsely connected externally.

- Can be overlapping or non-overlapping.
- Some networks may not have any meaningful community structure.
- Examples:
  - Social groups in social networks.
  - Proteins with similar functionality in protein interaction networks.



Faskowitz et al., 2018

A robust baseline method: spectral clustering

A generalized framework for clustering nodes of a network:

- simple,
- relatively robust,
- fast.

A robust baseline method: spectral clustering

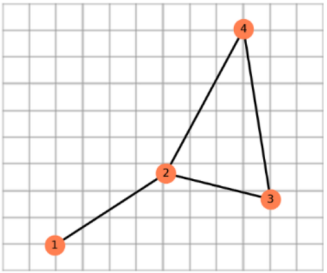
For a given $n \times n$ adjacency matrix A and a fixed number of communities K, spectral clustering procedure can be described as follows;

A robust baseline method: spectral clustering

For a given $n \times n$ adjacency matrix A and a fixed number of communities K, spectral clustering procedure can be described as follows;

1. Compute degree-normalized Laplacian matrix L by;

$$L = D^{-1/2} A D^{-1/2} \text{ where } D_{i,j} = \begin{cases} \deg(v_i) & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases}$$



Adjacency Matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

Normalized Laplacian Matrix

$$\mathcal{L} = \begin{pmatrix} 1 & -\frac{1}{\sqrt{3}} & 0 & 0 \\ -\frac{1}{\sqrt{3}} & 1 & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \\ 0 & -\frac{1}{\sqrt{6}} & 1 & -\frac{1}{2} \\ 0 & -\frac{1}{\sqrt{6}} & -\frac{1}{2} & 1 \end{pmatrix}$$

A robust baseline method: spectral clustering

For a given $n \times n$ adjacency matrix A and a fixed number of communities K, spectral clustering procedure can be described as follows;

1. Compute degree-normalized Laplacian matrix L by;

$$L = D^{-1/2} A D^{-1/2} \text{ where } D_{i,j} = \begin{cases} \deg(v_i) & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases}$$

2. Find K leading eigenvectors of L corresponding to K largest eigenvalues of L in absolute value.

A robust baseline method: spectral clustering

For a given $n \times n$ adjacency matrix A and a fixed number of communities K, spectral clustering procedure can be described as follows;

1. Compute degree-normalized Laplacian matrix L by;

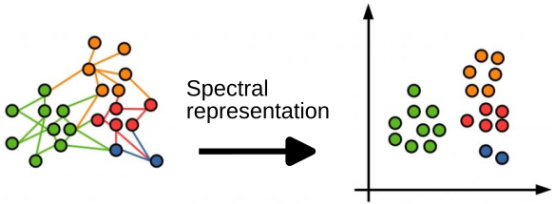
$$L = D^{-1/2} A D^{-1/2} \text{ where } D_{i,j} = \begin{cases} \deg(v_i) & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases}$$

2. Find K leading eigenvectors of L corresponding to K largest eigenvalues of L in absolute value.
3. Form the matrix $V \in \mathbb{R}^{n \times K}$ with the leading eigenvectors as columns.

A robust baseline method: spectral clustering

For a given $n \times n$ adjacency matrix A and a fixed number of communities K, spectral clustering procedure can be described as follows;

1. Compute degree-normalized Laplacian matrix L by;
2. Find K leading eigenvectors of L corresponding to K largest eigenvalues of L in absolute value.
3. Form the matrix $V \in \mathbb{R}^{n \times K}$ with the leading eigenvectors as columns.
4. Represent each node by its corresponding row in the matrix V, i.e. $v_u = V[u]$ corresponding to the spectral embedding of each node $u = 1, \ldots, n$.



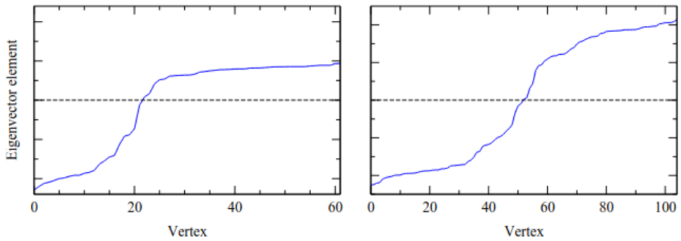
A robust baseline method: spectral clustering

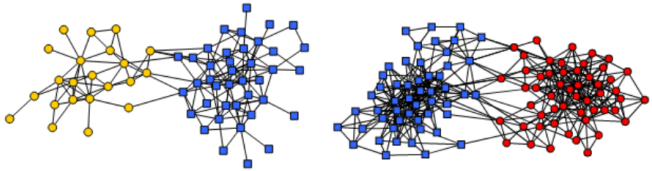
For a given $n \times n$ adjacency matrix A and a fixed number of communities K, spectral clustering procedure can be described as follows;

1. Compute degree-normalized Laplacian matrix L by;
2. Find K leading eigenvectors of L corresponding to K largest eigenvalues of L in absolute value.
3. Form the matrix $V \in \mathbb{R}^{n \times K}$ with the leading eigenvectors as columns.
4. Represent each node by its corresponding row in the matrix V, i.e. $v_u = V[u]$ corresponding to the spectral embedding of each node $u = 1, \ldots, n$.
5. Run chosen clustering algorithm (e.g. K-means) on the embeddings v_u, $u = 1, \ldots, n$.

A robust baseline method: spectral clustering

example





Dolphin social network

Political books

Newman, 2013.

Gene co-expression networks and gene modules

	S_1	S_2	S_3
G_1	43.26	40.89	5.05
G_2	166.6	41.87	136.65
G_3	12.53	39.55	42.09
G_4	28.77	191.92	236.56
G_5	114.7	79.7	99.76
G_6	119.1	80.57	114.59
G_7	118.9	156.69	186.95
G_8	3.76	2.48	136.78
G_9	32.73	11.99	118.8
G_{10}	17.46	56.11	21.41

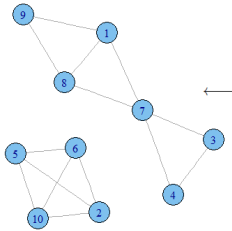
Gene expression values

$|r(G_i, G_j)|$

Pearson correlation

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	1.00	0.23	0.61	0.71	0.03	0.35	**0.86**	**1.00**	**0.97**	0.37
G_2	0.23	1.00	0.63	0.52	**0.98**	**0.99**	0.29	0.30	0.46	**0.99**
G_3	0.61	0.63	1.00	**0.99**	0.77	0.53	**0.93**	0.56	0.41	0.51
G_4	0.71	0.52	**0.99**	1.00	0.69	0.41	**0.97**	0.66	0.52	0.40
G_5	0.03	**0.98**	0.77	0.69	1.00	**0.95**	0.48	0.09	0.27	**0.94**
G_6	0.35	**0.99**	0.53	0.41	**0.95**	1.00	0.17	0.41	0.57	**1.00**
G_7	0.86	0.29	**0.93**	**0.97**	0.48	0.17	1.00	**0.83**	0.72	0.16
G_8	**1.00**	0.30	0.56	0.66	0.09	0.41	0.83	1.00	**0.98**	0.42
G_9	**0.97**	0.46	0.41	0.52	0.27	0.57	0.72	**0.98**	1.00	0.58
G_{10}	0.37	**0.99**	0.51	0.40	**0.94**	**1.00**	0.16	0.42	0.58	1.00

Similarity (Co-expression) score



	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	0	0	0	0	0	0	1	1	1	0
G_2	0	0	0	0	1	1	0	0	0	1
G_3	0	0	0	1	0	0	1	0	0	0
G_4	0	0	1	0	0	0	1	0	0	0
G_5	0	1	0	0	0	1	0	0	0	1
G_6	0	1	0	0	1	0	0	0	0	1
G_7	1	0	1	1	0	0	0	1	0	0
G_8	1	0	0	0	0	0	1	0	1	0
G_9	1	0	0	0	0	0	0	1	0	0
G_{10}	0	1	0	0	1	1	0	0	0	0

Network adjacency matrix

$|r(G_i, G_j)| >= 0.8$

Significance threshold

Gene co-expression networks and gene modules

	S_1	S_2	S_3
G_1	43.26	40.89	5.05
G_2	166.6	41.87	136.65
G_3	12.53	39.55	42.09
G_4	28.77	191.92	236.56
G_5	114.7	79.7	99.76
G_6	119.1	80.57	114.59
G_7	118.9	156.69	186.95
G_8	3.76	2.48	136.78
G_9	32.73	11.99	118.8
G_{10}	17.46	56.11	21.41

Gene expression values

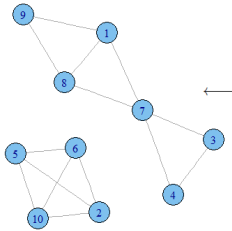
$|r(G_i, G_j)|$

Pearson correlation

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	1.00	0.23	0.61	0.71	0.03	0.35	**0.86**	**1.00**	**0.97**	0.37
G_2	0.23	1.00	0.63	0.52	**0.98**	**0.99**	0.29	0.30	0.46	**0.99**
G_3	0.61	0.63	1.00	**0.99**	0.77	0.53	**0.93**	0.56	0.41	0.51
G_4	0.71	0.52	**0.99**	1.00	0.69	0.41	**0.97**	0.66	0.52	0.40
G_5	0.03	**0.98**	0.77	0.69	1.00	**0.95**	0.48	0.09	0.27	**0.94**
G_6	0.35	**0.99**	0.53	0.41	**0.95**	1.00	0.17	0.41	0.57	**1.00**
G_7	0.86	0.29	**0.93**	**0.97**	0.48	0.17	1.00	**0.83**	0.72	0.16
G_8	**1.00**	0.30	0.56	0.66	0.09	0.41	0.83	1.00	**0.98**	0.42
G_9	**0.97**	0.46	0.41	0.52	0.27	0.57	0.72	**0.98**	1.00	0.58
G_{10}	0.37	**0.99**	0.51	0.40	**0.94**	**1.00**	0.16	0.42	0.58	1.00

Similarity (Co-expression) score

Edges encode co-expression strength between genes.



	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	0	0	0	0	0	0	1	1	1	0
G_2	0	0	0	0	1	1	0	0	0	1
G_3	0	0	0	1	0	0	1	0	0	0
G_4	0	0	1	0	0	0	1	0	0	0
G_5	0	1	0	0	0	1	0	0	0	1
G_6	0	1	0	0	1	0	0	0	0	1
G_7	1	0	1	1	0	0	0	1	0	0
G_8	1	0	0	0	0	0	1	0	1	0
G_9	1	0	0	0	0	0	0	1	0	0
G_{10}	0	1	0	0	1	1	0	0	0	0

Network adjacency matrix

$|r(G_i, G_j)| >= 0.8$

Significance threshold

Gene co-expression networks and gene modules

	S_1	S_2	S_3
G_1	43.26	40.89	5.05
G_2	166.6	41.87	136.65
G_3	12.53	39.55	42.09
G_4	28.77	191.92	236.56
G_5	114.7	79.7	99.76
G_6	119.1	80.57	114.59
G_7	118.9	156.69	186.95
G_8	3.76	2.48	136.78
G_9	32.73	11.99	118.8
G_{10}	17.46	56.11	21.41

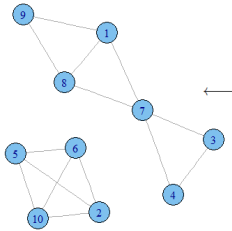
Gene expression values

$|r(G_i, G_j)|$

Pearson correlation

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	1.00	0.23	0.61	0.71	0.03	0.35	**0.86**	**1.00**	**0.97**	0.37
G_2	0.23	1.00	0.63	0.52	**0.98**	**0.99**	0.29	0.30	0.46	**0.99**
G_3	0.61	0.63	1.00	**0.99**	0.77	0.53	**0.93**	0.56	0.41	0.51
G_4	0.71	0.52	**0.99**	1.00	0.69	0.41	**0.97**	0.66	0.52	0.40
G_5	0.03	**0.98**	0.77	0.69	1.00	**0.95**	0.48	0.09	0.27	**0.94**
G_6	0.35	**0.99**	0.53	0.41	**0.95**	1.00	0.17	0.41	0.57	**1.00**
G_7	0.86	0.29	**0.93**	**0.97**	0.48	0.17	1.00	**0.83**	0.72	0.16
G_8	**1.00**	0.30	0.56	0.66	0.09	0.41	0.83	1.00	**0.98**	0.42
G_9	**0.97**	0.46	0.41	0.52	0.27	0.57	0.72	**0.98**	1.00	0.58
G_{10}	0.37	**0.99**	0.51	0.40	**0.94**	**1.00**	0.16	0.42	0.58	1.00

Similarity (Co-expression) score



	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	0	0	0	0	0	0	1	1	1	0
G_2	0	0	0	0	1	1	0	0	0	1
G_3	0	0	0	1	0	0	1	0	0	0
G_4	0	0	1	0	0	0	1	0	0	0
G_5	0	1	0	0	0	1	0	0	0	1
G_6	0	1	0	0	1	0	0	0	0	1
G_7	1	0	1	1	0	0	0	1	0	0
G_8	1	0	0	0	0	0	1	0	1	0
G_9	1	0	0	0	0	0	0	1	0	0
G_{10}	0	1	0	0	1	1	0	0	0	0

Network adjacency matrix

$|r(G_i, G_j)| >= 0.8$

Significance threshold

Edges encode co-expression strength between genes.

Community structure: strong local clustering of genes that are synchronized to function together.

scRNA-seq datasets across multiple individuals & time points

Question?

Can we construct networks from scRNA-seq data and perform community detection to find relevant gene modules?

scRNA-seq datasets across multiple individuals & time points

Question?

Can we construct networks from scRNA-seq data and perform community detection to find relevant gene modules?

Yes, you can. But you better be careful!

- Sparsity and noise cause missing and/or erroneous edges.
- Inferred community structure may not be reliable.

scRNA-seq datasets across multiple individuals & time points

Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation

Julie Jerber[1,2,10], Daniel D. Seaton[3,10], Anna S. E. Cuomo[3,10], Natsuhiko Kumasaka[2], James Haldane[2], Juliette Steer[2], Minal Patel[2], Daniel Pearce[2], Malin Andersson[2], Marc Jan Bonder[3], Ed Mountjoy[1], Maya Ghoussaini[1], Madeline A. Lancaster[4], HipSci Consortium*, John C. Marioni[2,3,5] ✉, Florian T. Merkle[6,7] ✉, Daniel J. Gaffney[2,11] ✉ and Oliver Stegle[2,3,8,9,11] ✉

Immune disease risk variants regulate gene expression dynamics during CD4+ T cell activation

Blagoje Soskic[1,2,5], Eddie Cano-Gamez[1,2,5], Deborah J. Smyth[1], Kirsty Ambridge[1], Ziying Ke[1], Julie C. Matte[1], Lara Bossini-Castillo[1], Joanna Kaplanis[1,2], Lucia Ramirez-Navarro[1], Anna Lorenc[1], Nikolina Nakic[3], Jorge Esparza-Gordillo[3], Wendy Rowan[3], David Wille[3], David F. Tough[3], Paola G. Bronson[4] and Gosia Trynka[1,2] ✉

Jerber-2021 Dataset

- scRNA-seq of iPS cells.
- 3 cell types, 2 time points, $\sim$ 20 donors.

Soskic-2022 Dataset

- scRNA-seq of CD4+ T cells.
- 2 cell types, 4 time points, $\sim$ 120 donors.

Defining problem motivated by multi-subject dynamic networks

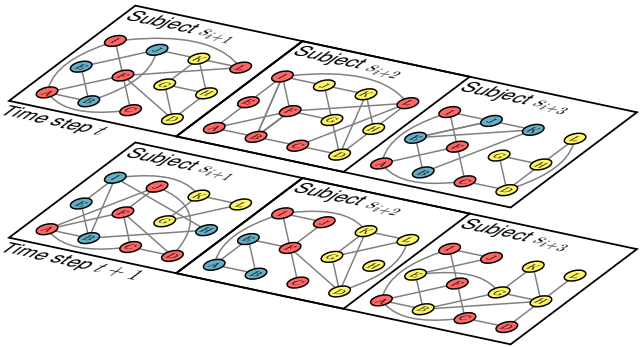
Problem

Given a multi-subject dynamic gene co-expression network, we aim to infer the *communities* for each time point and subject.

Defining problem motivated by multi-subject dynamic networks

Problem

Given a multi-subject dynamic gene co-expression network, we aim to infer the *communities* for each time point and subject.

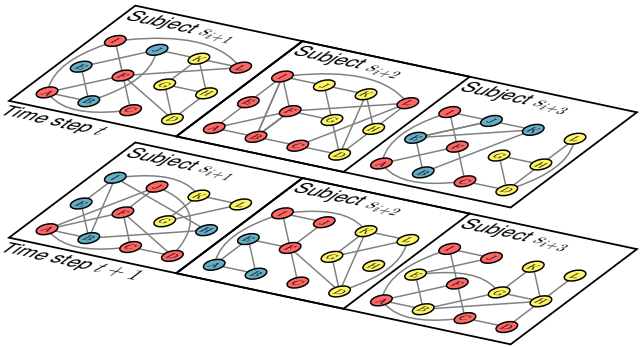


Time-series of adjacency matrices for each subject s:
$A_{s,1}, \ldots, A_{s,t} \ldots, A_{s,T}$.

Defining problem motivated by multi-subject dynamic networks

Problem

Given a multi-subject dynamic gene co-expression network, we aim to infer the *communities* for each time point and subject.



Time-series of adjacency matrices for each subject s:
$A_{s,1}, \ldots, A_{s,t} \ldots, A_{s,T}$.

We want to leverage signal shared across time and among subjects.

The goal is to infer meaningful communities from noisy networks!

Finding persistent communities by eigenvector smoothing

$V_{s,t}$: eigenvectors of degree normalized Laplacian of the adjacency matrix $A_{s,t}$.
$U_{s,t}$: projection matrix onto the column space of $V_{s,t}$, i.e., $U_{s,t} = V_{s,t} V_{s,t}^T$

Assumption of temporal smoothness of a single subject.

$$\min_{\substack{\overline{U}_{s,t} \\ t=1,\ldots,T}} \sum_{t=1}^{T} \left\| U_{s',t} - \overline{U}_{s',t} \right\|_F^2 + \sum_{t=1}^{T-1} \alpha \left\| \overline{U}_{s,t} - \overline{U}_{s,t+1} \right\|_F^2$$

$$\text{subject to } \overline{U}_{s',t}, \in \left\{ VV^T \ : \ V \in \mathbb{R}^{G \times K}, \ V^T V = I \right\} \ \forall t.$$

- $\alpha \left\| \overline{U}_{s',t} - \overline{U}_{s',t+1} \right\|_F^2$ enforces smoothness over the time dimension.

Promoting signal sharing among subjects

Assumption of similarity among subjects at a fixed time point.

$$\min_{\substack{\overline{U}_{s,t} \\ s=1,\dots,S \\ t=1,\dots,T}} \sum_{t=1}^{T} \left\| U_{s',t} - \overline{U}_{s',t} \right\|_F^2 + \sum_{t=1}^{T-1} \alpha \left\| \overline{U}_{s,t} - \overline{U}_{s,t+1} \right\|_F^2 + \sum_{t=1}^{T} \beta \left\| \overline{U}_{s,t} - \mu_s(\overline{U}_{:,t}) \right\|_F^2$$

$$\text{subject to } \overline{U}_{s',t}, \in \left\{ VV^T \ : \ V \in \mathbb{R}^{G \times K}, \ V^T V = I \right\} \ \forall s, \forall t.$$

- $\alpha \left\| \overline{U}_{s,t} - \overline{U}_{s,t+1} \right\|_F^2$ enforces smoothness over the time dimension.
- $\beta \left\| \overline{U}_{s,t} - \mu_s(\overline{U}_{:,t}) \right\|_F^2$ constrains subject-specific variations from the time-dependent mean projection matrix $\mu_s(\overline{U}_{:,t})$ and promotes signal sharing among subjects:

$$\mu_s(\overline{U}_{:,t}) = \frac{1}{S-1} \sum_{\substack{1 \le s' \le S \\ s' \neq s}} \overline{U}_{s',t}.$$

A simple yet effective iterative algorithm: MuDCoD

Global optimum of the optimization problem we defined is given by

$$
\begin{aligned}
\overline{U}_{s,1}^{\ell+1} &= \Pi_K\Big(U_{s,1} + \alpha \overline{U}_{s,2}^{\ell} + \beta\, \mu_s\Big(\overline{U}_{:,1}^{\ell}\Big)\Big) \\
\overline{U}_{s,t}^{\ell+1} &= \Pi_K\Big(\alpha \overline{U}_{s,t-1}^{\ell} + U_{s,t} + \alpha \overline{U}_{s,t+1}^{\ell} + \beta\, \mu_s\Big(\overline{U}_{:,t}^{\ell}\Big)\Big) \\
\overline{U}_{s,T}^{\ell+1} &= \Pi_K\Big(\alpha \overline{U}_{s,T-1}^{\ell} + U_{s,T} + \beta\, \mu_s\Big(\overline{U}_{:,T}^{\ell}\Big)\Big)\,,
\end{aligned}
\tag{1}
$$

where $t = 2, \ldots, T-1$, $s = 1, \ldots, S$ for all $\ell \geq 0$ and $\overline{U}_{s,t}^{0} = U_{s,t}$ for $\forall t, \forall s$.

The mapping $\Pi_K(M)$ extracts the K leading eigenvectors of the matrix M.

A simple yet effective iterative algorithm: MuDCoD

Global optimum of the optimization problem we defined is given by

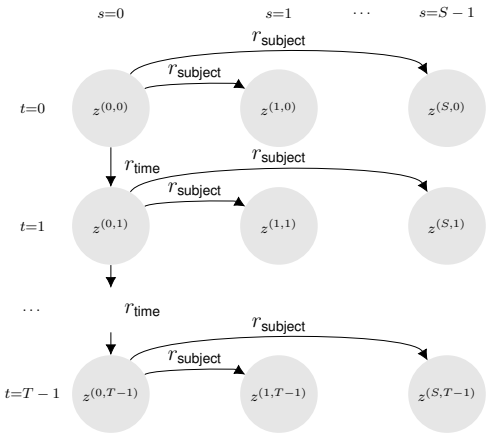
$$
\begin{aligned}
\overline{U}_{s,1}^{\ell+1} &= \Pi_K\Big(U_{s,1} + \alpha \overline{U}_{s,2}^{\ell} + \beta\, \mu_s\Big(\overline{U}_{:,1}^{\ell}\Big)\Big) \\
\overline{U}_{s,t}^{\ell+1} &= \Pi_K\Big(\alpha \overline{U}_{s,t-1}^{\ell} + U_{s,t} + \alpha \overline{U}_{s,t+1}^{\ell} + \beta\, \mu_s\Big(\overline{U}_{:,t}^{\ell}\Big)\Big) \\
\overline{U}_{s,T}^{\ell+1} &= \Pi_K\Big(\alpha \overline{U}_{s,T-1}^{\ell} + U_{s,T} + \beta\, \mu_s\Big(\overline{U}_{:,T}^{\ell}\Big)\Big),
\end{aligned}
\tag{1}
$$

where $t = 2, \ldots, T-1$, $s = 1, \ldots, S$ for all $\ell \geq 0$ and $\overline{U}_{s,t}^{0} = U_{s,t}$ for $\forall t, \forall s$.

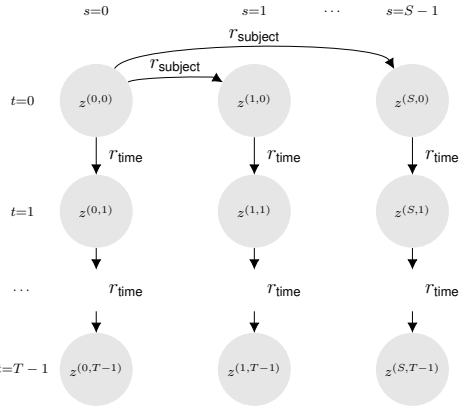
The mapping $\Pi_K(M)$ extracts the K leading eigenvectors of the matrix M.

- We allow K to vary over time and across subjects.
- For each network, after each iteration ℓ, K is inferred using eigenvalue statistics.

Evaluating the performance of MuDCoD on simulation data

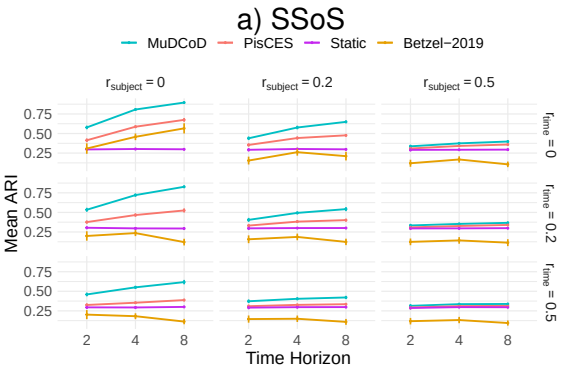


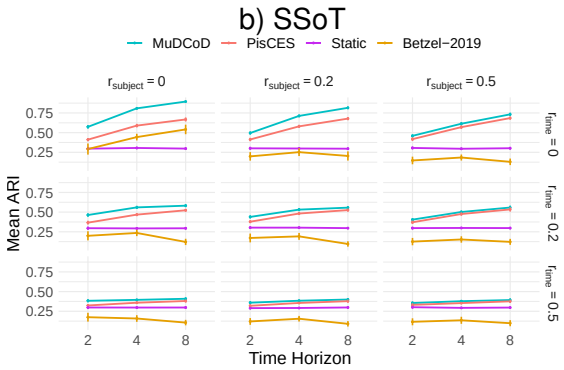
(a) *SSoS setting:* subjects evolve from a common ancestor at each time step t; only the ancestor's evolution over time is parameterized.



(b) *SSoT setting:* subjects evolve from a common ancestor at $t{=}0$; and then they evolve independently over time.

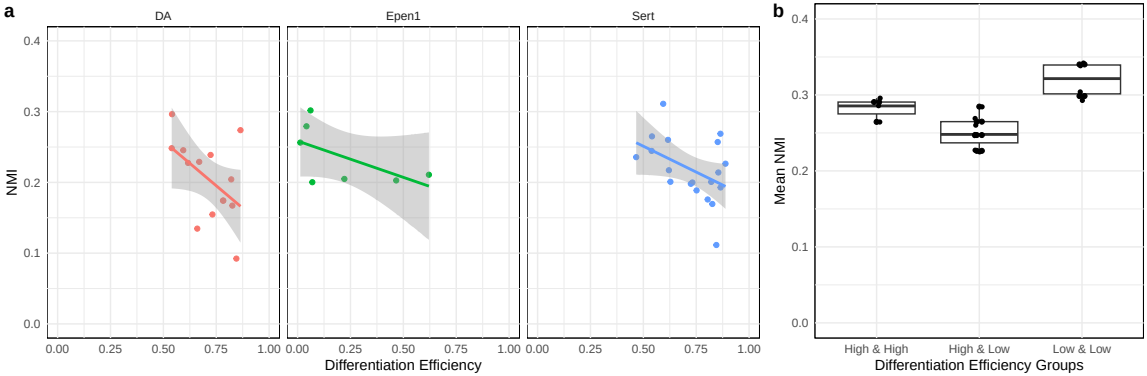
Evaluating the performance of MuDCoD on simulation data





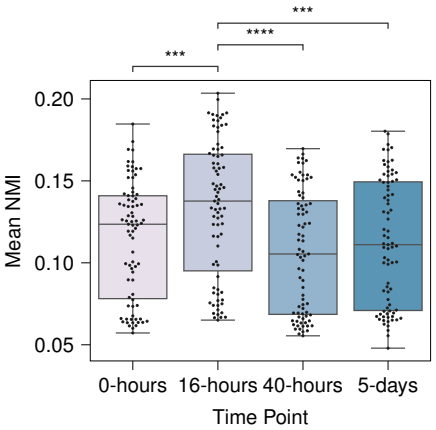
Applying MuDCoD to HipSci Consortium et al. [2021] data

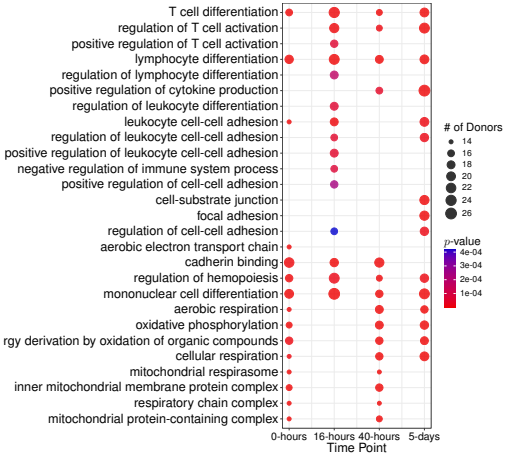
Communities found by MuDCoD can recover differentiation efficiency information.



Applying MuDCoD to Soskic et al. [2022] data

MuDCoD yields gene modules that associate with specific biological conditions.





Conclusion and future directions

Information sharing among subjects and along the time helps to infer relevant communities from noisy networks.

Conclusion and future directions

Information sharing among subjects and along the time helps to infer relevant communities from noisy networks.

Extending MuDCoD:

- Considering dissimilar subgroups of individuals present in the data, e.g., healthy and diseased.

Conclusion and future directions

Information sharing among subjects and along the time helps to infer relevant communities from noisy networks.

Extending MuDCoD:

- Considering dissimilar subgroups of individuals present in the data, e.g., healthy and diseased.
- Aligning communities inferred for different subjects, analyzing joint modules among subjects, flow between communities along the time.

Acknowledgment

Thank You!

Oznur Tastan, Sabanci University, Istanbul, Turkiye
Sunduz Keles, University of Wisconsin-Madison, Madison, WI, USA
Ferhat Ay, University of California San Diego, La Jolla, CA, USA

References

HipSci Consortium, Julie Jerber, Daniel D. Seaton, Anna S. E. Cuomo, Natsuhiko Kumasaka, James Haldane, Juliette Steer, Minal Patel, Daniel Pearce, Malin Andersson, Marc Jan Bonder, Ed Mountjoy, Maya Ghoussaini, Madeline A. Lancaster, John C. Marioni, Florian T. Merkle, Daniel J. Gaffney, and Oliver Stegle. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nature Genetics*, 53(3):304–312, March 2021. doi: 10.1038/s41588-021-00801-6. URL http://www.nature.com/articles/s41588-021-00801-6.

Blagoje Soskic, Eddie Cano-Gamez, Deborah J. Smyth, Kirsty Ambridge, Ziying Ke, Julie C. Matte, Lara Bossini-Castillo, Joanna Kaplanis, Lucia Ramirez-Navarro, Anna Lorenc, Nikolina Nakic, Jorge Esparza-Gordillo, Wendy Rowan, David Wille, David F. Tough, Paola G. Bronson, and Gosia Trynka. Immune disease risk variants regulate gene expression dynamics during CD4+ T cell activation. *Nature Genetics*, 54(6): 817–826, June 2022. ISSN 1546-1718. doi: 10.1038/s41588-022-01066-3. URL https://www.nature.com/articles/s41588-022-01066-3. Number: 6 Publisher: Nature