Spaceland Embedding of Sparse Stochastic Graphs

IEEE High Performance Extreme Computing
September 25, 2019




Nikos Pitsianis[12] Alexandros-S. Iliopoulos[2] Dimitris Floros[1] Xiaobai Sun[2]

[1]Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki
[2]Department of Computer Science, Duke University

Outline

1. **Introduction**
2. **Contribution A: SG-t-SNE**
3. **Contribution B: SG-t-SNE-Π**
4. **Key references**

1. Introduction
- **Graph embedding**
- **Precursor work**
- **Significant impact**
- **Main limitations**

2. Contribution A: SG-t-SNE

3. Contribution B: SG-t-SNE-Π

4. Key references

Introduction: graphs & graph embedding

Graph/network $\mathbf{G(V, E)}$: relational data

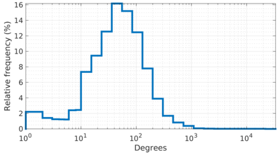
increasingly arise in various applications:

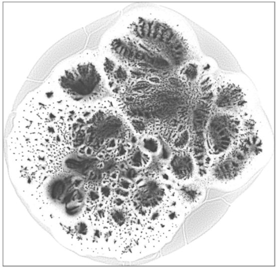
biological, social, friend networks, food webs, co-author networks, word co-occurrence networks, product co-purchase networks, . . .

Graph (vertex) embedding:
Mapping/encoding: $V = \mathcal{X} \implies \mathcal{Y} \subset \mathbb{R}^d$

- word embedding (of a co-occurrence graph)
- image embedding (of a nearest-similarity graph)
- product embedding (of a co-purchase graph)
- user embedding (of a friend network)

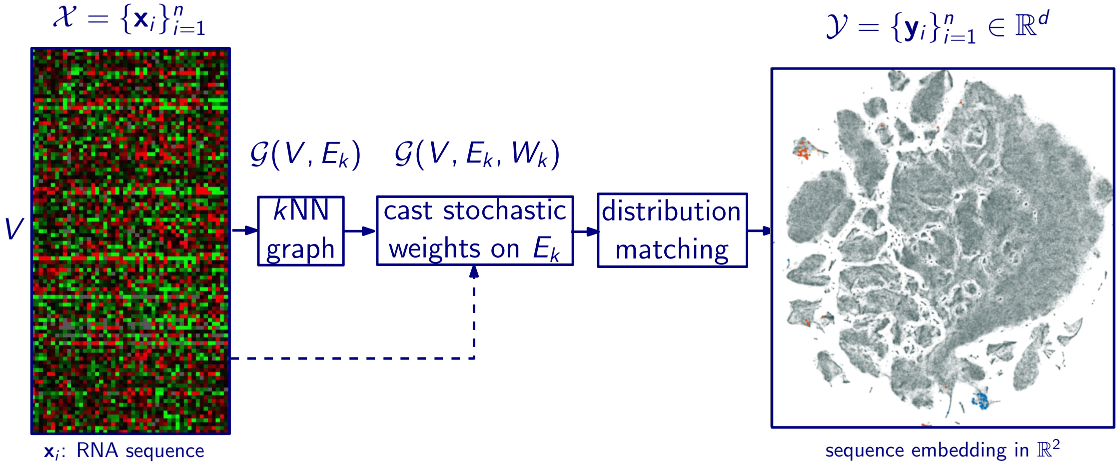
to facilitate many tasks of graph data analysis





Social network **orkut** with $\mathbf{n} = \mathbf{3{,}072{,}441}$ user nodes and $\mathbf{m} = \mathbf{237{,}442{,}607}$ friendship links: Degree distribution (top) and 2D embedding (bottom)

SNE: stochastic neighbor embedding algorithm



SNE[1] **pipeline** illustrated with spatial embedding of
$\mathbf{n} = \mathbf{1{,}306{,}127}$ RNA sequences of E18 mouse brain cells

[1]Hinton and Roweis, NIPS, 2003   10x Genomics, App Note, 2017

t-SNE: t-distributed SNE

From input vertex data $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$

Find kNNs among $\mathbf{D} = [d^2(x_i, x_j)]_{n\times n}$

Cast $\mathbf{D}_{\text{kNN}}$ to stochastic $\mathbf{P} = [p_{j|i} + p_{i|j}]/2$

$$p_{j|i}(\sigma_i) = \frac{1}{Z_i} \exp\left(-d_{ij}^2/2\sigma_i^2\right) \quad \text{(Gaussians)}$$

with σ_i determined by the perplexity equations

$$-\sum_j a_{ij} p_{j|i}(\sigma_i) \log(p_{j|i}(\sigma_i)) = \log(u), \quad \forall i \quad (1)$$

u: ***perplexity*** parameter chosen by the user

Vertex embedding coordinates

$$\mathcal{Y} = \{\mathbf{y}_i\}_{i=1}^n \in \mathbb{R}^d, \quad d = 1, 2, 3, \dots$$

Follow t-distribution (Cauchy kernel)

$$\mathbf{Q}: \quad q_{ij} = \frac{1}{Z}(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}$$

Determined by the best distribution matching measured by KL divergence[1]

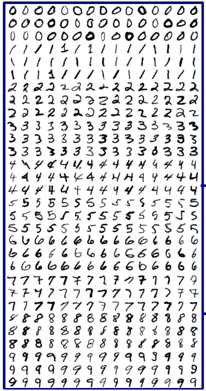
$$\mathcal{Y}^* = \arg\min_{\mathcal{Y}} \mathrm{KL}(\mathbf{P} \| \mathbf{Q}(\mathcal{Y}))$$

[1] van der Maaten and Hinton, JMLR, 2008

t-SNE: iterative embedding process

$\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$

V

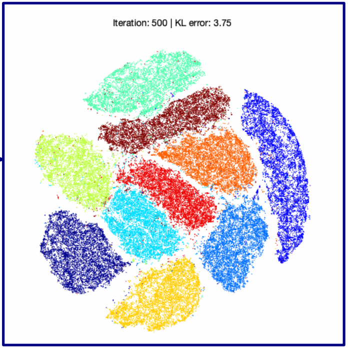


$\mathbf{x}_i$: pixels in digit image

$\mathcal{G}(V, E_k)$ $\mathcal{G}(V, E_k, W_k)$

kNN graph → cast stochastic weights on E_k → distribution matching →

$\mathcal{Y} = \{\mathbf{y}_i\}_{i=1}^n \in \mathbb{R}^d$



digit embedding in $\mathbb{R}^2$

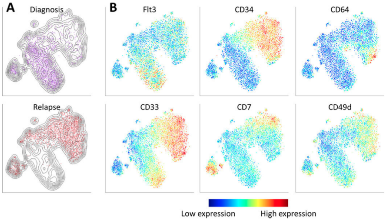
SNE[1] **pipeline** illustrated with spatial embedding of
$\mathbf{n} = \mathbf{60{,}000}$ handwritten digits (MNIST dataset)

[1] Hinton and Roweis, NIPS, 2003 LeCun et al., Proc IEEE, 1998

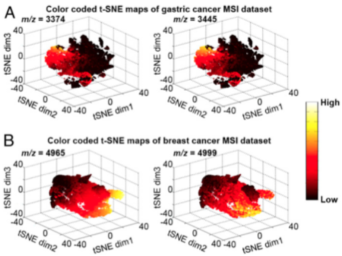
Significant impacts

With low-dim. spatial embedding in particular, the SNE/t-SNE algorithm family has enabled

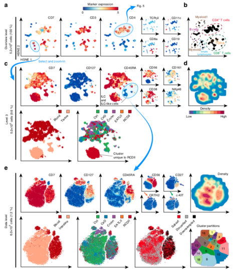
- visual inspection, identification of connections/separations
- network-based analysis for hidden connections
- hypothesis generating and scientific discoveries



Amir et al., Nat Biotechnol, 2013



Abdelmoula et al., PNAS, 2016



van Unen et al., Nat Commun, 2017

Main limitations

- Restricted to data in a metric space
- Restricted to kNN-based stochastic graphs

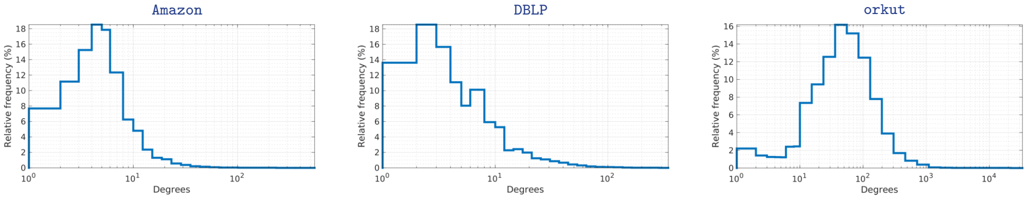
  Degree k and perplexity u are coupled by condition $\mathbf{0} < \mathbf{u} < \mathbf{k}$ implied in (1)

Vertices of a network do not necessarily readily reside in a metric space

A typical economic phenomenon:

low-degree nodes in majority
hub nodes in minority

Irregular in degree distribution
Defying the parameter condition $\mathbf{u} < \deg(\mathbf{i})$



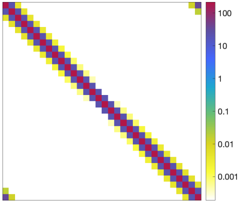
Irregular degree distribution for each of three real-world networks:
Low-degree nodes (including leaf nodes) in majority; high-degree nodes in minority.

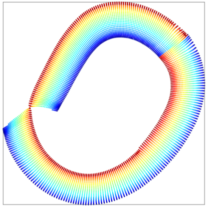
Main limitations

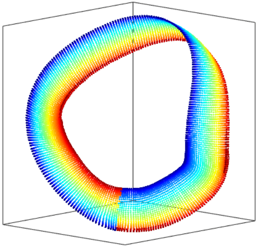
- Existing software programs* are limited, due to slow computation speed, to
  - small graphs, or
  - 1D/2D embedding

Many networks are large;

Spaceland (3D) embedding has much greater potential in preserving/encoding more structural information



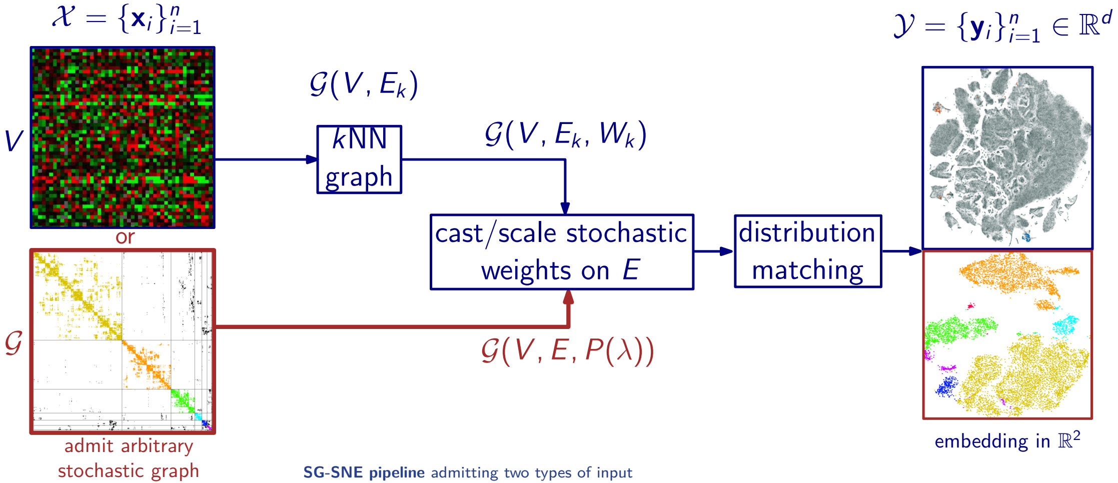




(Left) kNN graph ($k = 150$) for a Möbius strip on a 256×32 lattice, with $n = 8{,}192$ nodes,
(Middle) 2D embedding with missed/unresolved connections,
(Right) 3D embedding with correct connections, also offering multiple or steerable views.

* van der Maaten, JMLR, 2014 `https://lvdmaaten.github.io/tsne` Linderman et al., Nat Methods, 2019 `https://github.com/KlugerLab/FIt-SNE`

SG-t-SNE: stochastic graph t-SNE



SG-SNE pipeline admitting two types of input
(top) embedding of $\mathbf{n} = \mathbf{1{,}306{,}127}$ RNA sequences of E18 mouse brain cells
(bottom) embedding of $\mathbf{n} = \mathbf{8{,}381}$ peripheral blood mononuclear cells

10x Genomics, App Note, 2017 Zheng et al., Nat Commun, 2017

SG-t-SNE: distinctive extension & the keystone

Distinctions:

- ◇ Admitting arbitrary stochastic graph $\mathbf{P} = [p_{j|i}]$
  i.e., extend the embedding to the entire family of stochastic graphs
- ◇ Making it feasible to exploit sparse connection pattern for
  - investigative/explorative data analysis
  - higher computation efficiency

Key: the stochastic reshaping/rescaling equations: $\forall i$

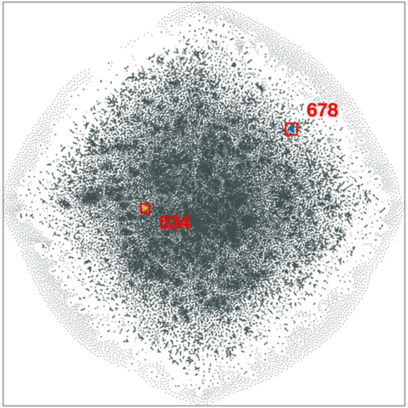
$$\sum_j a_{ij}\,\phi\left(p_{j|i}^{\gamma_i}\right) = \lambda \qquad \Longrightarrow \qquad p_{j|i}(\lambda) = \frac{a_{ij}\,\phi\left(p_{j|i}^{\gamma_i}\right)}{\lambda},$$

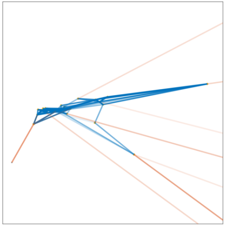
$\boldsymbol{\lambda} > 0$: re-scaling parameter; $\phi \geq 0$: reshaping function, monotonically increasing [1]

$\mathbf{A} = [a_{ij}]$: the binary-valued adjacency matrix; Solutions $\boldsymbol{\gamma_i}$ exist unconditionally

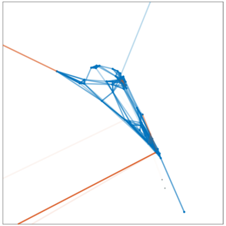
[1] We used $\phi(x) = x$ for the presented embeddings

Enabled embedding of Amazon product co-purchase network





(534)



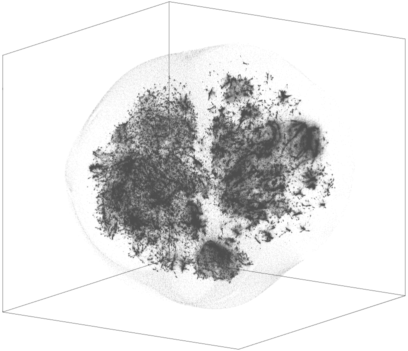
(678)

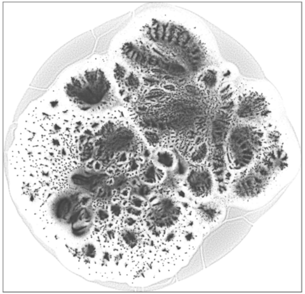
ID	n_{sub}	e_{in}	e_{out}	w_{in}	w_{out}
(534)	44	374	20	71.7	2.4
(678)	70	506	19	114.6	3.3

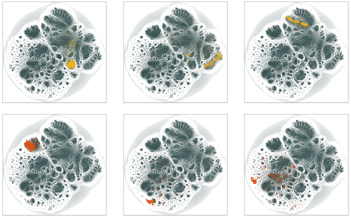
Amazon product sale network: $\boldsymbol{n} = \mathbf{334{,}863}$ products, $\boldsymbol{m} = \mathbf{1{,}851{,}744}$ edges for co-purchase connectivity, irregular degree distribution. (Left) 2D product embedding enabled by SG-t-SNE; (Right) two product clusters/subgraphs, the vertices for each are embedded closer together, with denser intra-connections.

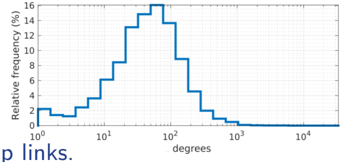
Yang and Leskovec, K&IS, 2015

Enabled embedding of social network **orkut**









Social network **orkut**: $\boldsymbol{n} = \mathbf{3{,}072{,}441}$ user nodes, $\boldsymbol{m} = \mathbf{237{,}442{,}607}$ friendship links.

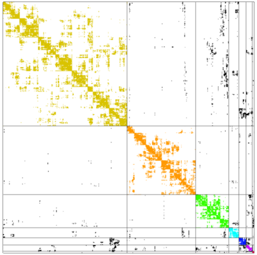
(Left & Middle) 3D and 2D embeddings enabled by SG-t-SNE;

(Right) **Findings**: There is a weak-link zone (easier to observe in 3D embedding), calibrated communities reside on one or the other side; the rich structure reflects/decodes information of geophysical regions and cultural diversities.

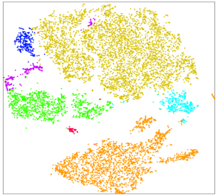
Yang and Leskovec, K&IS, 2015

SG-t-SNE: exploiting sparse patterns

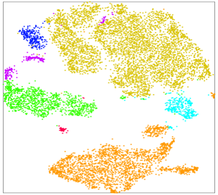
- ▷ Vertex data: $8k$ peripheral blood mononuclear cells (PBMCs)
- ▷ PBMC embedding via kNN graphs by a cell similarity measure
- ▷ SG-t-SNE can use a much sparser neighbor graph



kNN graph $\mathbf{P}_k$, $k = 30$



t-SNE: $\boldsymbol{k}=\mathbf{150}$, $u=50$

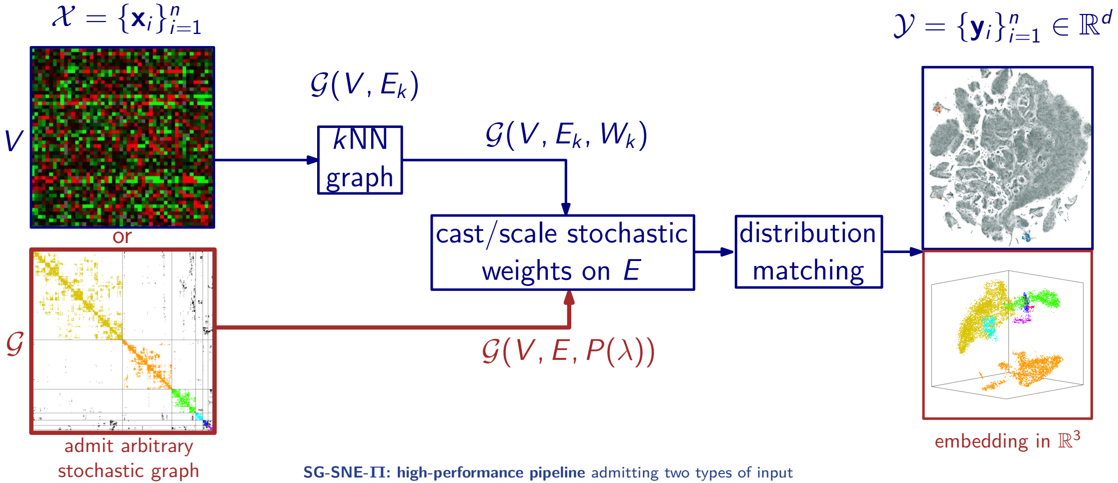


SG-t-SNE: $\boldsymbol{k}=\mathbf{30}$, $\lambda=80$

PBM cells are color coded by provided labels with the data.

Zheng et al., Nat Commun, 2017

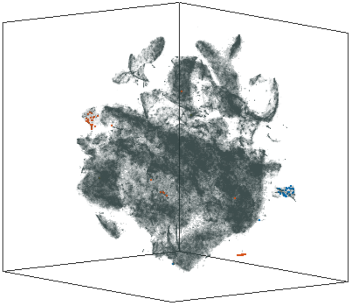
SG-t-SNE-Π: enabling spaceland (3D) embedding

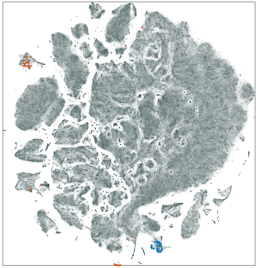


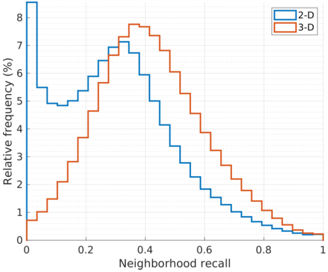
SG-SNE-Π: high-performance pipeline admitting two types of input
(top) embedding of $\mathbf{n} = \mathbf{1{,}306{,}127}$ RNA sequences of E18 mouse brain cells
(bottom) embedding of $\mathbf{n} = \mathbf{8{,}381}$ peripheral blood mononuclear cells

10x Genomics, App Note, 2017 Zheng et al., Nat Commun, 2017

Comparison in neighborhood preservation







(Left and Middle) RNA sequence embeddings in 3D and 2D, respectively, via kNN graph ($k = 90$) with $\boldsymbol{n} = \mathbf{1{,}306{,}127}$ RNA sequences of E18 mouse brain cells and 1,000 principle gene components.
(Right) Comparison in neighborhood recall shows the advantage of 3D embedding

10x Genomics, App Note, 2017

Iterative embedding search: computational challenges

The computation bulk is in iterative gradient updates.
van der Maaten re-formulated the gradient in two interaction terms

$$\frac{\partial(\,\mathrm{KL}(\mathbf{P}||\mathbf{Q}(\mathcal{Y}))\,)}{\partial \mathbf{y}_i} = \frac{4}{Z} \underbrace{\sum_{i \neq j} p_{ij} q_{ij} (\mathbf{y}_i - \mathbf{y}_j)}_{\text{attractive interaction}} - \frac{4}{Z} \underbrace{\sum_{i \neq j} q_{ij}^2 (\mathbf{y}_i - \mathbf{y}_j)}_{\text{repulsive interaction}},$$

- ▷ **PQ** $= [p_{ij} q_{ij}]$: kernel matrix of the attraction term, **sparse** and **irregular**
- ▷ **QQ** $= [q_{ij} q_{ij}]$: kernel matrix of the repulsion term, **full** and **irregular**

  with **exploitable structure** by Barnes-Hut algorithm[1] or by nuFFT-based factorization[2]
- ▷ Both sparse and compressive interactions tend to suffer from memory latency or inadequate parallel scheduling due to irregular memory accesses

  Each term in need of high-performance algorithm-software support, especially on desktop, laptop computers for individual researchers

[1] van der Maaten, JMLR, 2014 [2] Linderman et al., Nat Methods, 2019

Accelerated gradient updates by SG-t-SNE-Π

$$\frac{\partial(\mathrm{KL})}{\partial \mathbf{y}_i} = \frac{4}{Z} \underbrace{\sum_{i \neq j} p_{ij} q_{ij} (\mathbf{y}_i - \mathbf{y}_j)}_{\text{attraction: } \mathbf{PQ} = [p_{ij} q_{ij}]} - \frac{4}{Z} \underbrace{\sum_{i \neq j} q_{ij}^2 (\mathbf{y}_i - \mathbf{y}_j)}_{\text{repulsion: } \mathbf{QQ} = [q_{ij} q_{ij}]},$$

Fast interaction with sparse PQ

- same sparse pattern as **P**, which is reordered (once) to a pattern of block sparse with denser blocks (BSDB)
- modified Compressed Sparse Blocks (CSB) library[1]

Fast interaction with compressed QQ

- utilized an internal equi-spaced grid in two ways
- scattered data points binned into grid cells
- formulated a kernel splitting on the grid instead of augmenting the grid size by $2x$ in each dimension

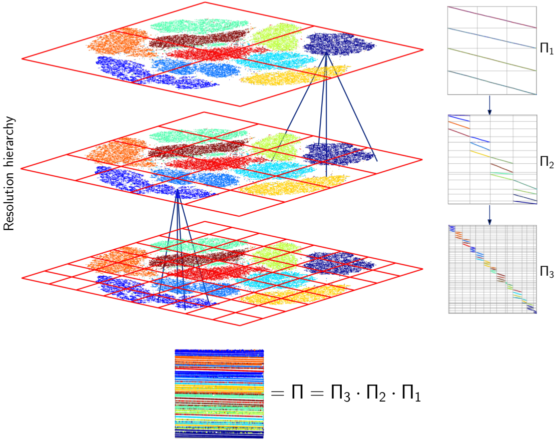
[1] Buluç et al., ASPAA, 2009 Pitsianis et al., JOSS, 2019

Multi-level data translocation in SG-t-SNE-Π

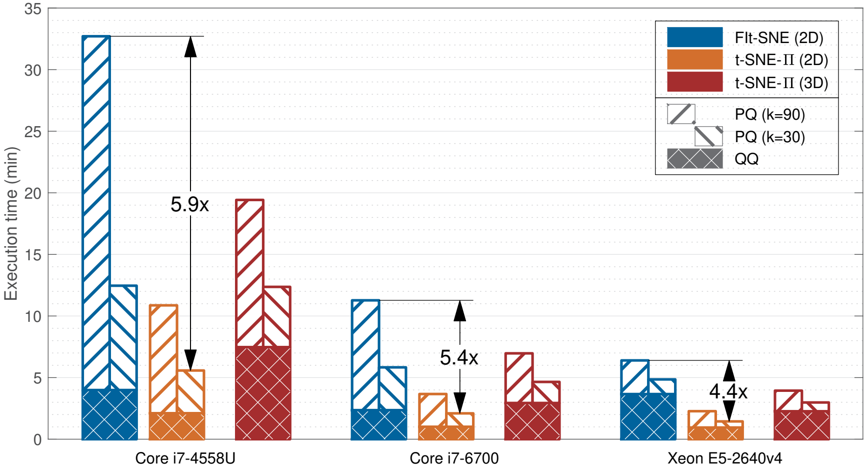
- By $\boldsymbol{\Pi}$ we refer to **data permutation and physical relocations** within each interaction, also in between, at every iteration step
- The **fast data translocation problem**

  Data $\mathcal{Y}$ available in ordering $a(\mathcal{Y})$, to be accessed in a different ordering $b(\mathcal{Y})$

  Determine a data translocation scheme to carry out $\Pi : a(\mathcal{Y}) \rightarrow b(\mathcal{Y})$ in shortest time subject to computing platform specifics
- **Solution**: architecture-adaptive decomposition of the permutation toward
  - optimal data locality
  - maximal utilization of parallel resources
  - best payoff with data translocation overhead



Comparison in execution time



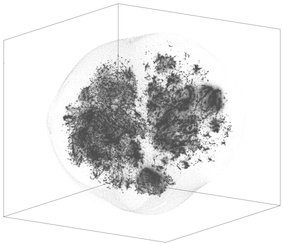
Comparison in execution time for embedding of kNN graphs, $k \in \{30, 90\}$, with $n = 1{,}306{,}127$ nodes as single-cell RNA sequences of E18 mouse brain cells.[1] Each embedding takes 1,000 iterations and maintains an approximation error below the same tolerance (10^{-6}).

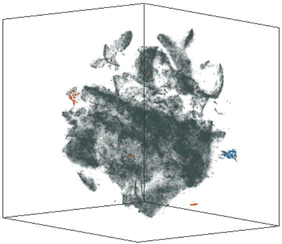
[1] 10x Genomics, App Note, 2017

Recap

- ▷ **SG-t-SNE**: enables embedding of arbitrary stochastic graphs
  - including kNN graphs generated by vertex data
  - embeddings of large real-world graphs reveal characteristic structures and new information

- ▷ **SG-t-SNE-Π** enables fast spaceland (3D) embedding
  - preserve more neighborhood connection, structure info.
  - offer multiple vantage points
  - open source software and supplementary material at http://t-sne-pi.cs.duke.edu





Acknowledgments: We thank Tiancheng Liu and anonymous reviewers for their valuable input and comments, thank Pantazis Vouzaxakis and Xenofon Theodoridis for their assitance in experiments.

References I

10x Genomics. Transcriptional profiling of 1.3 million brain cells with the Chromium single cell 3' solution. *LIT000015 Chromium Million Brain Cells Application Note*, 2017.

W. M. Abdelmoula, B. Balluff, S. Englert, J. Dijkstra, M. J. T. Reinders, A. Walch, L. A. McDonnell, and B. P. F. Lelieveldt. Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data. *Proceedings of the National Academy of Sciences*, 113 (43):12244–12249, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1510227113.

E. D. Amir, K. L. Davis, M. D. Tadmor, E. F. Simonds, J. H. Levine, S. C. Bendall, D. K. Shenfeld, S. Krishnaswamy, G. P. Nolan, and D. Pe'er. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology*, 31(6):545–552, 2013. doi: 10.1038/nbt.2594.

A. Buluç, J. T. Fineman, M. Frigo, J. R. Gilbert, and C. E. Leiserson. Parallel sparse matrix-vector and matrix-transpose-vector multiplication using compressed sparse blocks. In *Proceedings of Annual Symposium on Parallelism in Algorithms and Architectures*, pages 233–244, 2009. doi: 10.1145/1583991.1584053.

G. E. Hinton and S. T. Roweis. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*, volume 15, pages 857–864, 2003.

Y. LeCun, L. Bottou, B. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

G. C. Linderman, M. Rachh, J. G. Hoskins, S. Steinerberger, and Y. Kluger. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature Methods*, 16(3):243–245, 2019. doi: 10.1038/s41592-018-0308-4.

N. Pitsianis, D. Floros, A.-S. Iliopoulos, and X. Sun. SG-t-SNE-Π: Swift neighbor embedding of sparse stochastic graphs. *Journal of Open Source Software*, 4(39): 1577, 2019. ISSN 2475-9066. doi: 10.21105/joss.01577.

L. van der Maaten. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, 15:3221–3245, 2014.

L. van der Maaten and G. E. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

V. van Unen, T. Höllt, N. Pezzotti, N. Li, M. J. T. Reinders, E. Eisemann, F. Koning, A. Vilanova, and B. P. F. Lelieveldt. Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. *Nature Communications*, 8(1):1740, 2017. doi: 10.1038/s41467-017-01689-9.

J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213, 2015. doi: 10.1007/s10115-013-0693-z.

G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017. doi: 10.1038/ncomms14049.