Pattern Recognition Letters 128 (2019) 100–106

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Doubly Stochastic Neighbor Embedding on Spheres ☆

Yao Lu [a,b], Jukka Corander [c,d], Zhirong Yang [a,e,*]

[a] Department of Computer Science, Aalto University, Finland
[b] College of Engineering and Computer Science, Australian National University, Australia
[c] Department of Department of Biostatistics, University of Oslo, Norway
[d] Department of Mathematics and Statistics, University of Helsinki, Finland
[e] Department of Computer Science, Norwegian University of Science and Technology, Norway

ARTICLE INFO

Article history:
Received 24 August 2018
Revised 21 August 2019
Accepted 26 August 2019
Available online 26 August 2019

Keywords:
Data visualization
Nonlinear dimensionality reduction
Information divergence

ABSTRACT

Stochastic Neighbor Embedding (SNE) methods minimize the divergence between the similarity matrix of a high-dimensional data set and its counterpart from a low-dimensional embedding, leading to widely applied tools for data visualization. Despite their popularity, the current SNE methods experience a crowding problem when the data include highly imbalanced similarities. This implies that the data points with higher total similarity tend to get crowded around the display center. To solve this problem, we introduce a fast normalization method and normalize the similarity matrix to be doubly stochastic such that all the data points have equal total similarities. Furthermore, we show empirically and theoretically that the doubly stochasticity constraint often leads to embeddings which are approximately spherical. This suggests replacing a flat space with spheres as the embedding space. The spherical embedding eliminates the discrepancy between the center and the periphery in visualization, which efficiently resolves the crowding problem. We compared the proposed method (DOSNES) with the state-of-the-art SNE method on three real-world datasets and the results clearly indicate that our method is more favorable in terms of visualization quality. DOSNES is freely available at http://yaolubrain.github.io/dosnes/.

1. Introduction

Information visualization by dimensionality reduction facilitates a viewer to quickly digest information in massive data. It is therefore increasingly applied as a critical component in scientific research, digital libraries, data mining, financial data analysis, market studies, manufacturing production control and drug discovery, etc. Numerous dimensionality reduction methods have been introduced, ranging from linear methods such as Principal Component Analysis to nonlinear methods such as Multidimensional Scaling (MDS), [MDS; 14], Isomap [13], Locally Linear Embedding [10], Gaussian Process Latent Variable Models [6]. A survey on nonlinear dimensionality reduction has been given by van der Maaten et al. [9]. Aspects in Multidimensional Scaling are discussed by Buja et al. [1].

Recently, Stochastic Neighbor Embedding (SNE) and its variants [4,8,12] have achieved remarkable progress in data visualization, especially for displaying clusters in data. An SNE method takes as input the pairwise similarities between data points in the high-dimensional space and tries to preserve the similarities in a low-dimensional space by minimizing the Kullback–Leibler divergence between the input and output similarity matrices.

The input to SNE is a similarity matrix or the affinity matrix of a weighted graph. When the node degrees of the graph are highly imbalanced, SNE tends to place the high-degree nodes in the center and the low-degree ones in the periphery, regardless of the intrinsic similarities between the nodes. Therefore, SNE often experiences the “crowding-in-the-center” problem for highly imbalanced affinity graphs.

We propose two techniques to overcome the above-mentioned drawback. First, we impose a doubly stochasticity constraint on the input similarity matrix. Two-way normalization has been shown to improve spectral clustering [16] and here we verify that it is also beneficial for data visualization. Moreover, if the neighborhood graph is asymmetric, for example, k-Nearest-Neighbors (kNN) or entropy affinities [8,15], we provide an efficient method for converting it to a doubly stochastic matrix.

Second, we observe that the data points are often distributed approximately around a sphere if the input similarity matrix is

☆ Handling by Associate Editor: Kar-Ann Toh.

* Corresponding author at: Department of Computer Science, Norwegian University of Science and Technology, Norway.
E-mail address: zhirong.yang@ntnu.no (Z. Yang).

https://doi.org/10.1016/j.patrec.2019.08.026

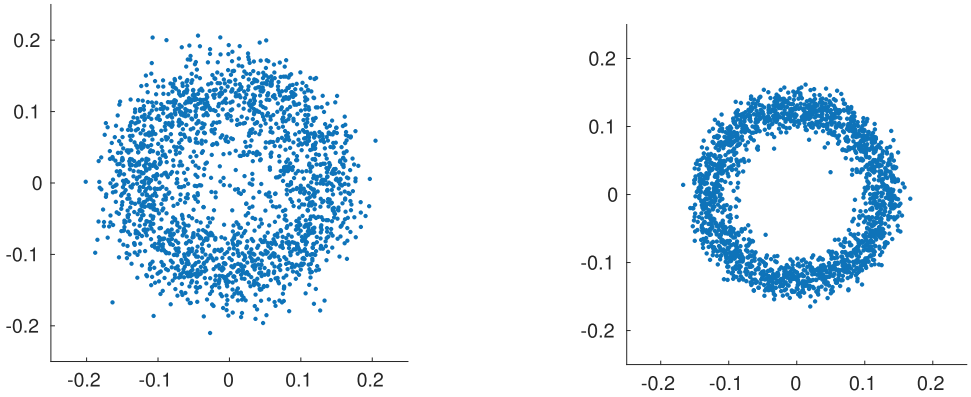


Fig. 1. t-SNE visualization of a random uniformly distributed matrix (left) and a random doubly stochastic matrix (right).

doubly stochastic, and we provide a theoretical analysis of this phenomenon. Our analysis suggests replacing the two-dimensional Euclidean embedding space with spheres in the three-dimensional space. Since there is no global center or periphery on the sphere geometry, the visualization is then naturally free of "crowding-in-the-center" problem. Moreover, we present an efficient projection step for adapting an SNE method with the spherical constraint.

We tested the proposed method on several real-world datasets and compared it with the state-of-the-art SNE method, t-SNE [8]. The new method is superior to t-SNE in resolving the crowding problem and in preserving intrinsic similarities.

In the next section we briefly review SNE methods. We then discuss doubly stochastic similarity matrix and spherical embedding in Sections 3 and 4, respectively. We present experimental results in Section 5 and conclusions in Section 6.

2. Stochastic Neighbor Embedding

Stochastic Neighbor Embedding [SNE; 4] is a nonlinear dimensionality reduction method. Given a set of multivariate data points $\{x_1, x_2, \ldots, x_n\}$, where $x_i \in \mathbb{R}^D$, their neighborhood is encoded in a square nonnegative matrix P, where P_{ij} is the probability that x_j is a neighbor of x_i. SNE finds a mapping $x_i \mapsto y_i \in \mathbb{R}^d$ for $i = 1, \ldots, n$ such that the neighborhoods are approximately preserved in the mapped space. Usually the mapping is defined such that $d = 2$ or 3, and $d < D$. If the neighborhood in the mapped space is encoded in $Q \in \mathbb{R}^{n\times n}$ such that Q_{ij} is the probability that y_j is a neighbor of y_i, the SNE task is to minimize the Kullback–Leibler divergence $\mathcal{D}_{\mathrm{KL}}(P||Q)$ over $Y = [y_1, y_2, \ldots, y_n]^T$.

Symmetric Stochastic Neighbor Embedding [s-SNE; 8] is a variant of SNE. Given input similarity $p_{ij} \geq 0$, s-SNE minimizes Kullback-Leibler divergence between the matrix-wise normalized similarities $P_{ij} = p_{ij}/\sum_{ab} p_{ab}$ and $Q_{ij} = q_{ij}/\sum_{ab} q_{ab}$. The output similarity q_{ij} is typically chosen to be proportional to a Gaussian distribution so that $q_{ij} = \exp\left(-\|y_i - y_j\|^2\right)$, or proportional to a Cauchy distribution so that $q_{ij} = (1 + \|y_i - y_j\|^2)^{-1}$. The Cauchy s-SNE method is also called t-Distributed Stochastic Neighbor Embedding [t-SNE; 8]. The optimization of s-SNE can be implemented with the gradients for Gaussian case: $\partial\mathcal{J}/\partial y_i = 4\sum_j (P_{ij} - Q_{ij})(y_i - y_j)$ and for Cauchy case $\partial\mathcal{J}/\partial y_i = 4\sum_j (P_{ij} - Q_{ij})(y_i - y_j)q_{ij}$. Here $4\sum_j P_{ij}(y_i - y_j)$ or $4\sum_j P_{ij}(y_i - y_j)q_{ij}$ can be interpreted as the attractive force for y_i, while $-4\sum_j Q_{ij}(y_i - y_j)$ or $-4\sum_j Q_{ij}(y_i - y_j)q_{ij}$ as the repulsive force.

3. Doubly stochastic similarity matrix

The input to s-SNE, P, is a nonnegative and symmetric matrix and can be treated as the affinity matrix of an undirected weighted graph. If the degree (i.e., row sum or column sum of P) distribution of nodes is highly non-uniform, then the high-degree nodes will usually receive and emit more attractive force than the average nodes during the iterative learning. As a result, these nodes often glue together and form the center of display. On the other hand, the low-degree nodes tend to be placed in the periphery due to less attraction. This behavior is often undesired in visualization because it only reveals the data centrality but hinders the discovery of other useful patterns, and may be directly misleading when some high-degree nodes are actually disconnected in the underlying data.

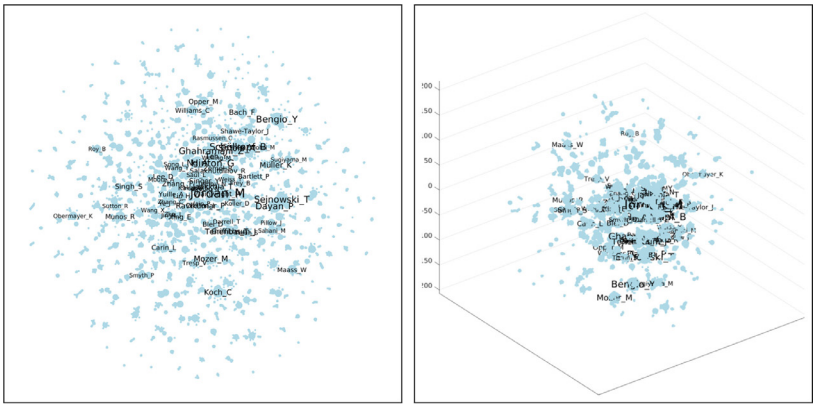
To overcome the above drawback, we can normalize the graph affinity such that the nodes have the same degree. For undirected graphs, this can be implemented by replacing the unitary matrix-wise sum constraint $\sum_{ij} P_{ij} = 1$ in s-SNE with the doubly stochasticity constraint, i.e., $\sum_i P_{ij} = \sum_j P_{ij} = 1$.

Given a non-normalized matrix, we can apply Sinkhorn–Knopp [11] or Zass-Shashua method [16] to project it to the closest doubly stochastic matrix P. In this work we use the former because it can maintain the sparsity of in the similarity matrix, which is often needed for large-scale tasks. Given a non-normalized similarity matrix S, the Sinkhorn–Knopp method initializes $P = S$ and iterates the following update rules until P has converged: for all i, $u_i \leftarrow \Sigma_j P_{ij}$, and then for all i, j, $P_{ij} \leftarrow P_{ij} u_i^{-1/2} u_j^{-1/2}$.

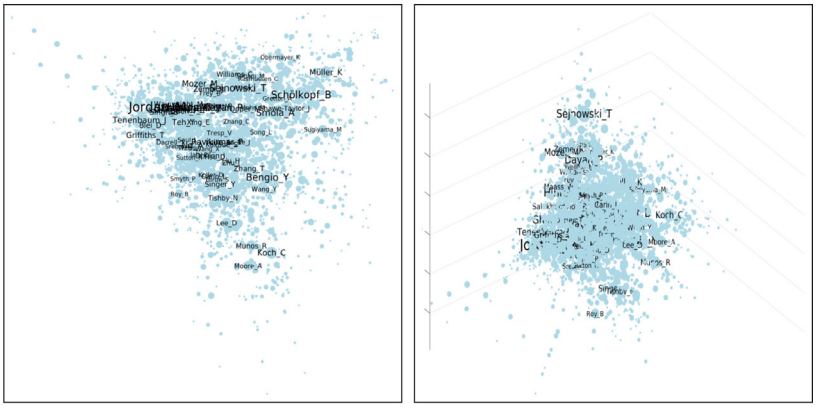
Alternatively, the neighborhood information in high-dimensional space can be encoded in an asymmetric matrix $B \geq 0$ with n rows, for example, the kNN graph or the entropy affinities [8,15]. B can also be a non-square dyadic data such as document-term or author-paper co-occurrence matrix. In these cases, we can apply the following steps to construct a doubly stochastic matrix: suppose $\Sigma_k B_{ik} > 0$ for all i, we first calculate for all i, k, $A_{ik} \leftarrow B_{ik}/\Sigma_u B_{iu}$, and then for all i, j $P_{ij} \leftarrow \sum_k \frac{A_{ik}A_{jk}}{\sum_v A_{vk}}$. It is easy to verify that by this construction P is symmetric and doubly stochastic. The calculations of A and P are performed only once and are thus computationally much more efficient than Sinkhorn–Knopp method which needs iterative steps. Here the matrix A_{ik} can be treated as the random walk probability from the ith row index to the kth column index and P_{ij} is interpreted as the two-step random walk probability between two row indices i and j via any column index k (with uniform prior over row indices).

4. Spherical embedding of doubly stochastic similarity matrices

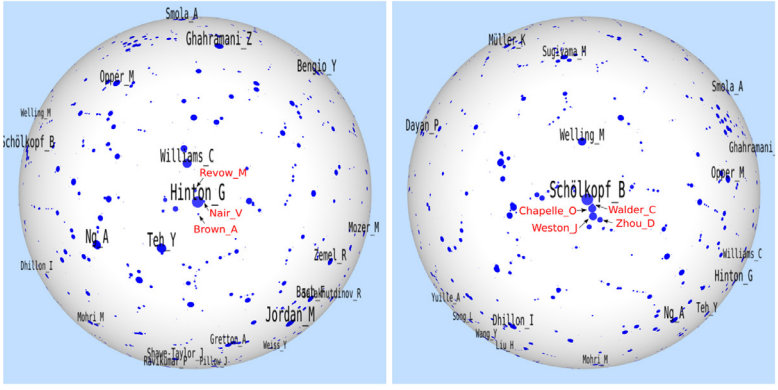
When the input similarity matrix is doubly stochastic, we find that s-SNE often embeds the data points around a sphere in the low-dimensional space. The phenomenon is illustrated in Fig. 1, where we generated a 2000 × 2000 similarity matrix with



(a) t-SNE 2D (b) t-SNE 3D



(c) MDS 2D (d) MDS 3D



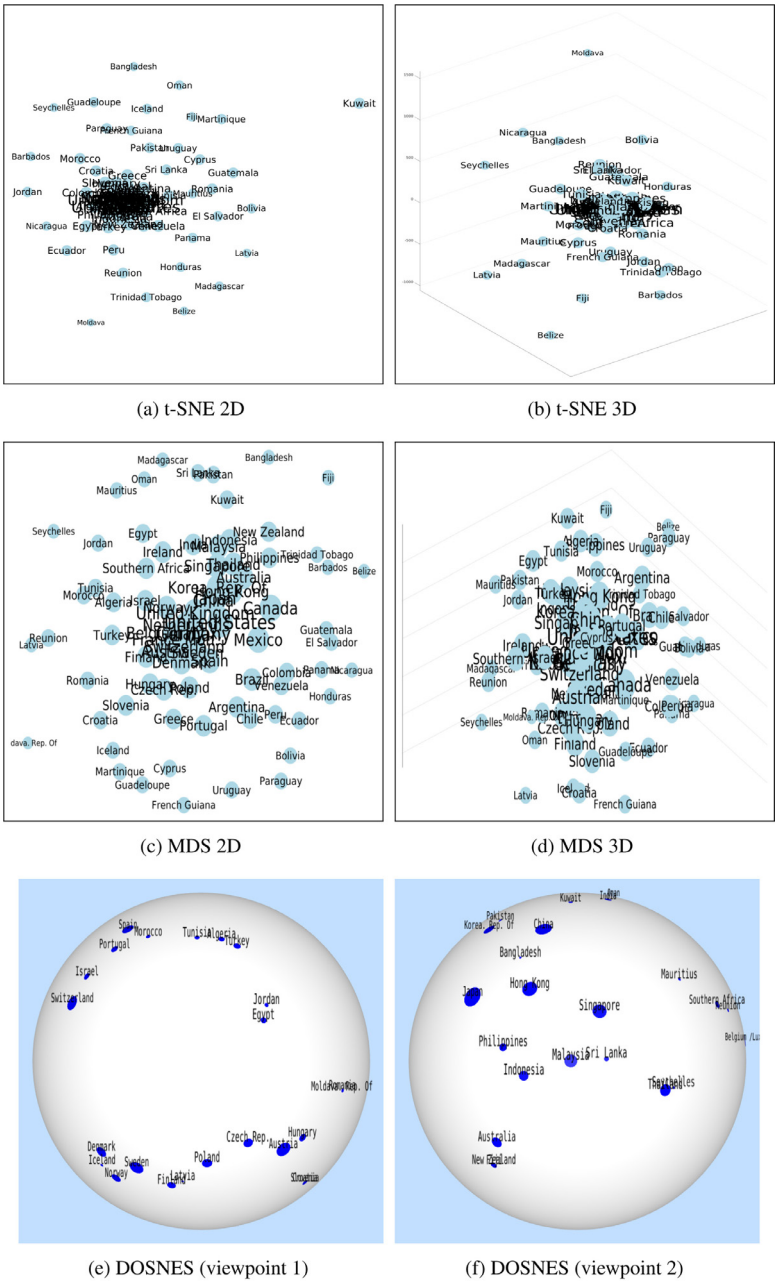
(e) DOSNES (viewpoint 1) (f) DOSNES (viewpoint 2)

Fig. 2. Visualizations of the NIPS dataset.

uniform distribution and visualize it by t-SNE. We can see from the left subfigure that the embedding is close to a ball. In contrast, if the matrix is doubly stochastically normalized (by using the Sinkhorn–Knopp method), the resulting embedded points approximately lie around a circle. The same phenomenon also holds for 3D visualizations.

We provide a theoretical analysis of this phenomenon. If P is doubly stochastic, then Q is often approximately doubly stochastic (up to a constant factor) because it approximates P by the KL-divergence. That is, $\Sigma_j Q_{ij}$ is approximately the same for all i. For example, in Fig. 1 (right), $\Sigma_j Q_{ij}$ mainly distribute around a constant (with mean 0.0005 and very small standard deviation 1.7×10^{-6}). In this case, we show that $\sum_j \|y_i - y_j\|^2$ becomes approximately the same for all i, bounded by constants, in Proposition 4.1. Furthermore, we show that when $\sum_j \|y_i - y_j\|^2$ is exactly the same for all i, the embedded points must be on a sphere, in Proposition 4.2. The proofs of the propositions are provided in the supplemental document.

Proposition 4.1. *If* $\sum_j q_{ij} = c$ *for* $i = 1, \ldots, n$ *and* $c > 0$, *then* $L \le \sum_j \|y_i - y_j\|^2 \le U$, *where (1) for* $q_{ij} = \exp(-\|y_i - y_j\|^2)$, $L = n \ln \frac{n}{c}$ *and* $U = n \ln \frac{n}{c - nb}$, *with* $b = a + (1-a)m - m^a$, $m = \min_j \exp(-\|y_i - y_j\|^2)$ *and* $a = \dfrac{\ln[\ln(1/m)/(1-m)]}{\ln(1/m)}$; *(2) for* $q_{ij} = (1 + \|y_i - y_j\|^2)^{-1}$, $L = \frac{n^2}{c} - n$ *and* $U = \frac{n^2}{c} - n + n(\sqrt{b} - 1)^2$, *with* $b = 1 + \max_j \|y_i - y_j\|^2$.



(a) t-SNE 2D (b) t-SNE 3D

(c) MDS 2D (d) MDS 3D

(e) DOSNES (viewpoint 1) (f) DOSNES (viewpoint 2)

Fig. 3. Visualizations of the `WorldTrade` dataset.

Proposition 4.2. *If* $\sum_j \|y_i - y_j\|^2 = c$ *for* $i = 1, \ldots, n$, $c > 0$ *and* $\sum_i y_i = 0$, *then* $\|y_1\|^2 = \|y_2\|^2 = \cdots = \|y_n\|^2$.

The propositions show that embeddings are often nearly spherical for doubly stochastic similarity matrices. Therefore it is more suitable to replace the 2D Euclidean embedding space with spheres in 3D space. The resulting layout can be encoded with $n \times 2 + 1$ numbers (two angles for each data point plus the common radius). Therefore the embedding is still intrinsically two-dimensional.

The spherical geometry itself brings other benefits for visualization. First, the embedding in the Euclidean space has a global center in the middle, while on spheres there is no such global center. Therefore a spherical visualization is free of the "crowding-in-the-center" problem. Every point on the sphere can be a local center, which provides fish-eye views for navigation and for examining patterns beyond centrality. Second, the attractive and repulsive forces can be transmitted in a cyclic manner, which helps in discovering macro patterns such inter-cluster similarities.

We thus formulate our learning objective as follows:

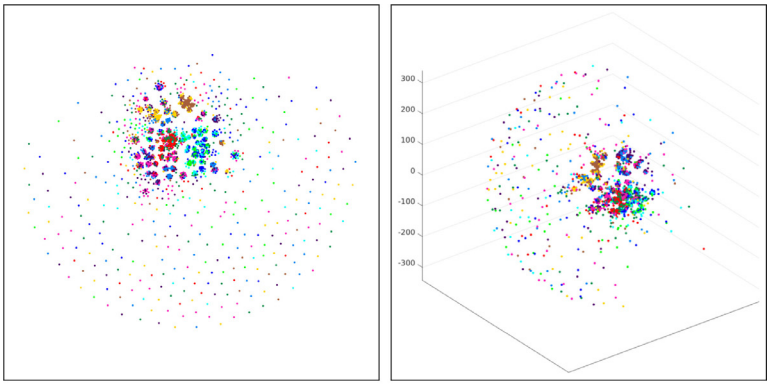
$$\underset{Y \in \mathbb{S}}{\text{minimize}} \quad \mathcal{J}(Y) = \mathcal{D}_{\text{KL}}(P||Q), \tag{1}$$

where $\mathcal{J}(Y)$ is an SNE objective function with P doubly stochastic, Q defined in Section 2 and

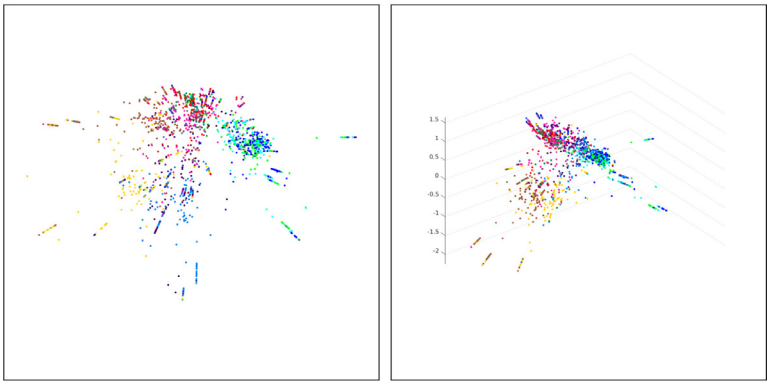
$$\mathbb{S} = \left\{ Y \;\middle|\; Y \in \mathbb{R}^{n \times 3};\ \|y_1\| = \cdots = \|y_n\|;\ \sum_i y_i = 0 \right\}. \tag{2}$$

We call the new method Doubly Stochastic Neighbor Embedding on Spheres (DOSNES).

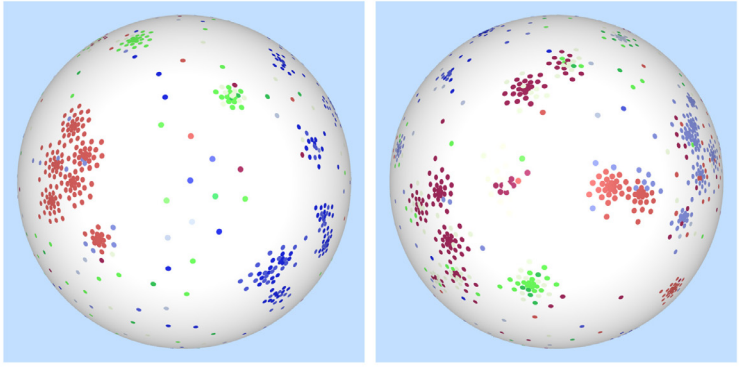
It is important to notice that our formulation is more flexible than other works on spherical embeddings (e.g., [2,3,7]). In DOSNES, the solution space $\mathbb{S}$ includes all centered spheres in the three-dimensional space, not only the sphere with unit or



(a) t-SNE 2D (b) t-SNE 3D



(c) MDS 2D (d) MDS 3D



(e) DOSNES (viewpoint 1) (f) DOSNES (viewpoint 2)

Fig. 4. Visualizations of the MIREX dataset.

pre-fixed radius. Moreover, we do not require normalization of the input vectors. Detailed comparison with related work is given in Section 2 of the supplemental document.

We employ a projection step after each SNE update step to enforce the sphere constraint. The DOSNES algorithm steps are summarized as follows:

1. Normalize P to be doubly stochastic.
2. Repeat until convergence
   (a) $\widetilde{Y} \leftarrow$OneStepUpdateSNE(P, Y),
   (b) $Y \leftarrow \arg\min_{Z\in\mathbb{S}} \|Z - \widetilde{Y}\|$.

The projection step 2b is performed by implicitly switching $\widetilde{Y} = [\tilde{y}_1, \ldots, \tilde{y}_n]^T$ to the spherical coordinate system, taking the mean radius, and switching back to Cartesian coordinates. This is implemented as: For $i = 1, \ldots, n$

$$y_i \leftarrow \frac{\tilde{y}_i}{\|\tilde{y}_i\|} \cdot \left(\frac{1}{n} \sum_j \|\tilde{y}_j\| \right), \tag{3}$$

where $\tilde{y}_i = \tilde{y}_i - \frac{1}{n}\sum_j \tilde{y}_j$. The iterations converge to a stationary point with suitable learning step sizes [see e.g., 5, Section 5].

5. Experiments

We developed a browser-based software for displaying and navigating the DOSNES results. The software and its demos can be found in the project website.[1] In the paper we present the 2D projected views of the spheres.

[1] http://yaolubrain.github.io/dosnes/

Table 1
Quantitative comparison: (top) K-means clustering purity and (bottom) running time (in seconds).

	DOSNES	t-SNE
WorldTrade	0.64	0.44
MIREX	0.40	0.31
	DOSNES	t-SNE
WorldTrade	0.1 s	0.1 s
MIREX	108.4 s	107.1 s
NIPS	333.3 s	328.9 s

We compare our proposed method DOSNES with two- and three-dimensional t-SNE[2] as well as non-metric MDS[3] in Euclidean embedding space [8] to verify the effectiveness of using doubly stochastic similarities and the sphere constraint.

The compared methods were tested on three real-world datasets from different domains:

(1) NIPS:[4] the proceedings of NIPS conferences (1987–2015) which contains 5993 papers and their associated 6621 authors. We used the largest connected component in the co-author graph with 5300 papers and 5422 authors. The (non-normalized) similarity matrix is from the co-author graph, i.e., BB^T where B is the author-paper co-occurrence matrix.

(2) WorldTrade:[5] trade network of metal manufactures among 80 countries in 1994. Each edge represents the total trade amount (imports and exports) between two countries.

(3) MIREX:[6] the dataset is from the Third Music Information Retrieval Evaluation eXchange (MIREX 2007). It is a network of 3090 songs in 10 music genre classes. The weighted edges are human judgment on how similar two songs are.

MDS requires a distance matrix as input. Given a similarity matrix S, we first normalize $\widetilde{S}_{ij} = S_{ij}/\max(S)$. Treating $\widetilde{S}_{ij}$s as cosine similarities, we obtain the cosine distances by $D_{ij} = 1 - \widetilde{S}_{ij}$. Next we calculate the shortest graph distances between all nodes and feed them to MDS.

The NIPS co-author graph is visualized in Fig. 2. The node degrees of the graph are highly uneven, where many authors have only one paper while the most productive author has 93 papers. In Fig. 2(a) and (b), we can see both 2D and 3D t-SNE caused the most productive NIPS authors crowded in the center. This is undesirable because these authors actually do not often co-author NIPS papers. For example, Hinton_G has no co-authored paper with Schölkopf_B but they are very close in the t-SNE layout. A similar crowding problem is observed in the MDS visualizations. In Fig. 2(e) and (f), DOSNES resolves neatly the crowding problem, by normalizing the similarity matrix with our method in Section 3 and visualizing the authors with spherical layout. The productive NIPS authors are now more evenly distributed. For example, Hinton_G becomes more distant to Schölkopf_B. Meanwhile, retrieval around the most established authors reveals accurate co-authorship. For example, Revow_M, Nair_V and Brown_A are close to Hinton_G because all their NIPS papers are co-authored with Hinton_G. See our online demo[7] for more details.

The visualizations of the WorldTrade graph are given in Fig. 3. In this graph, some countries such as United States and Germany have more total trade amount than many others. In Fig. 3(a)–(d), we can see both 2D and 3D t-SNE, as well as the MDS visualizations, caused these countries crowded in the center. In contrast, DOSNES places the countries more evenly. In Fig. 3(e) and (f), we can see on the sphere many meaningful clusters (e.g., Europe and Asia) which well match the geography even though we did not use such information in the training. See our demo globe[8] for other viewpoints.

Fig. 4 gives the visualizations of the MIREX dataset. In the panels (a) and (b), we can see that t-SNE caused over 90% of songs crowded in the center. A similar crowding problem appears in the MDS visualizations (panels c and d). In contrast, DOSNES performs much better in terms of separating the song genres and their subgroups, as in Fig. 4(e) and (f).

The effectiveness of DOSNES can be quantified by using the WorldTrade and MIREX data sets where ground truth classes are available. We performed K-means clustering on the DOSNES and t-SNE embeddings. The resulting cluster purities are reported in Table 1 (top). We can see that DOSNES achieves significantly higher purity for both data sets.

We also recorded the running time of DOSNES and t-SNE for the data sets. See Table 1 (bottom). DOSNES requires almost the same time as t-SNE, which shows that DOSNES improves t-SNE at negligible additional cost.

6. Conclusions

We have presented a new visualization method for high-dimensional and graph data. The proposed DOSNES method is based on the Stochastic Neighbor Embedding principle but with two key improvements: we normalize the input similarity matrix to be doubly stochastic and replace the 2D Euclidean embedding space with spheres in 3D space. Empirical results show that our method significantly outperforms the state-of-the-art approach t-SNE in terms of resolving the crowding problem and preserving intrinsic similarities.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patrec.2019.08.026.

References

[1] A. Buja, D.F. Swayne, M. Littman, N. Dean, H. Hofmann, L. Chen, Data visualization with multidimensional scaling, J. Comput. Graph. Stat. 17 (2) (2008) 444–472.

[2] https://lvdmaaten.github.io/tsne/
[3] We used the isoMDS() function in the MASS R package.
[4] https://papers.nips.cc/
[5] http://vlado.fmf.uni-lj.si/pub/networks/data/esna/metalWT.htm
[6] http://www.music-ir.org/mirex/wiki/2007
[7] http://yaolubrain.github.io/dosnes/demo/nips/
[8] http://yaolubrain.github.io/dosnes/demo/worldtrade/

[2] J. Deng, J. Guo, N. Xue, S. Zafeiriou, ArcFace: additive angular margin loss for deep face recognition, in: Proceedings of the CVPR, 2019.
[3] Y. Duan, J. Lu, J. Zhou, UniformFace: learning deep equidistributed representation for face recognition, in: Proceedings of the CVPR, 2019.
[4] G. Hinton, S. Roweis, Stochastic neighbor embedding, in: Proceedings of the NIPS, 2002.
[5] A. Iusem, On the convergence properties of the projected gradient method for convex optimization, Comput. Appl. Math. 22 (1) (2003) 37–52.
[6] N. Lawrence, Gaussian process latent variable models for visualisation of high dimensional data, in: Proceedings of the NIPS, 2004.
[7] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, SphereFace: deep hypersphere embedding for face recognition, in: Proceedings of the CVPR, 2017.
[8] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, JMLR 9 (2008) 2579–2605.
[9] L. van der Maaten, E. Postma, J. van den Herik, Dimensionality reduction: a comparative review, Technical Report, Tilburg University, TiCC TR, 2009-005.
[10] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[11] R. Sinkhorn, P. Knopp, Concerning nonnegative matrices and doubly stochastic matrices, Pac. J. Math. 21 (2) (1967) 343–348.
[12] J. Tang, J. Liu, M. Zhang, Q. Mei, Visualizing large-scale and high-dimensional data, in: Proceedings of the WWW, 2016.
[13] J. Tenenbaum, V. de Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[14] W. Torgerson, Multidimensional scaling: I. theory and method, Psychometrika 17 (4) (1952) 401–419.
[15] M. Vladymyrov, M. Carreira-Perpiñán, Entropic affinities: properties and efficient numerical computation, in: Proceedings of the ICML, 2013.
[16] R. Zass, A. Shashua, Doubly stochastic normalization for spectral clustering, in: Proceedings of the NIPS, 2006.