Mach Learn (2012) 87:33–55
DOI 10.1007/s10994-011-5273-4

Visualizing non-metric similarities in multiple maps

Laurens van der Maaten · Geoffrey Hinton

Received: 25 October 2010 / Accepted: 17 November 2011 / Published online: 17 December 2011

Abstract Techniques for multidimensional scaling visualize objects as points in a low-dimensional metric map. As a result, the visualizations are subject to the fundamental limitations of metric spaces. These limitations prevent multidimensional scaling from faithfully representing non-metric similarity data such as word associations or event co-occurrences. In particular, multidimensional scaling cannot faithfully represent intransitive pairwise similarities in a visualization, and it cannot faithfully visualize "central" objects. In this paper, we present an extension of a recently proposed multidimensional scaling technique called t-SNE. The extension aims to address the problems of traditional multidimensional scaling techniques when these techniques are used to visualize non-metric similarities. The new technique, called multiple maps t-SNE, alleviates these problems by constructing a collection of maps that reveal complementary structure in the similarity data. We apply multiple maps t-SNE to a large data set of word association data and to a data set of NIPS co-authorships, demonstrating its ability to successfully visualize non-metric similarities.

Keywords Multidimensional scaling · Embedding · Data visualization · Non-metric similarities

Editor: Paolo Frasconi.

L. van der Maaten (✉)
Pattern Recognition and Bioinformatics Laboratory, Delft University of Technology, Mekelweg 4, Delft 2628 CD, The Netherlands
e-mail: lvdmaaten@gmail.com

G. Hinton
Department of Computer Science, University of Toronto, 6 King's College Road, M5S 3G4 Toronto, ON, Canada
e-mail: hinton@cs.toronto.edu

1 Introduction

Classical scaling (Torgerson 1952) and other techniques for multidimensional scaling (e.g., Sammon 1969; Kruskal and Wish 1986; Tenenbaum et al. 2000; Belkin and Niyogi 2002; Schölkopf and Smola 2002; Cayton and Dasgupta 2006; Lafon and Lee 2006) represent similar objects, for instance, words that exhibit a certain semantic similarity, by nearby points in a low-dimensional metric map. Over the last decade, research on multidimensional scaling has focused on the development of more sophisticated similarity measurements between objects using, for instance, geodesic or diffusion distances (Tenenbaum et al. 2000; Lafon and Lee 2006) or kernels (Belkin and Niyogi 2002; Schölkopf et al. 1998; Schölkopf and Smola 2002). Another line of research has focused on learning similarity measurements between objects before performing classical scaling (Weinberger et al. 2005; Globerson and Roweis 2007; Shaw and Jebara 2009; Lawrence 2011). However, these approaches do not address the fundamental limitations of multidimensional scaling that are due to the characteristics of metric spaces (Tversky and Hutchinson 1986; Griffiths et al. 2007; Jäkel et al. 2008). A metric space is a space in which the following four *metric axioms* hold: (1) non-negativity of distances, (2) identity of indiscernibles, (3) symmetry of distances, and (4) the triangle inequality. If we denote the distance between object A and object B by $d(A, B)$, the four metric axioms may be denoted by

$$\begin{aligned} &d(A, B) \geq 0, \\ &d(A, B) = 0 \quad \text{iff} \quad A = B, \\ &d(A, B) = d(B, A), \\ &d(A, C) \leq d(A, B) + d(B, C). \end{aligned}$$

The metric axioms give rise to limitations of metric spaces in terms of the similarities that can be represented in these spaces. We mention two such limitations: (1) the triangle inequality that holds in metric spaces induces transitivity of similarities and (2) the number of points that can have the same point as their nearest neighbor is limited.[1] As a result of these limitations, multidimensional scaling cannot faithfully visualize similarity data that does not obey the metric axioms in a low-dimensional visualization. The aim of this paper is to construct visualizations that are not hampered by the two main limitations of metric spaces. We first discuss the two limitations in more detail below, using visualization of semantic similarities as an example.

The first limitation of metric spaces is due to the triangle inequality, which basically states that if point A is close to point B and B is close to point C, A has to be close to C as well. In practice, this constraint may well be violated by the implicit structure of similarity data. Consider, for instance, the word *tie*, which has a semantic similarity to words such as *suit* and *tuxedo*. In a low-dimensional metric map of the input objects, these three words need to be close to each other. However, the word *tie* is ambiguous: it is also semantically similar to words such as *rope* and *knot*, and should therefore be close to these words as well. As a result, the words *suit* and *rope* will be shown close together in a low-dimensional map of words even though the words exhibit very little similarity other than their association with *tie*.

[1] This is not the only limitation on the neighborhood relations of points in a metric space. For instance, the maximum number of equidistant points in a metric space is limited as well.

The second limitation of low-dimensional metric maps is that only a limited number of points can have the same point as their nearest neighbor. For instance, in a two-dimensional space, at most five points can have the same point as their nearest neighbor (by arranging them in a pentagon that is centered on the point). As a result, a low-dimensional metric map constructed by multidimensional scaling cannot faithfully visualize the large number of similarities of "central" objects with other objects. Similarity data may well contain such "central" objects. For instance, word meanings are characterized by a high "centrality", i.e., by the presence of words that are similar to a large portion of the other words (Tversky and Hutchinson 1986; Steyvers and Tenenbaum 2005). For instance, large numbers of mammals have a closer semantic similarity with the word *mammal* than with each other, and as a result, these mammals would like to have the word *mammal* as their nearest neighbor in a visualization. It is impossible to achieve this in a low-dimensional metric map of words because only a limited number of points can have the same nearest neighbor.

The reader should note that, although we use word similarities as an example of non-metric similarities throughout the paper, the same problems occur when visualizing many other types of similarity data. Prominent examples of non-metric similarity data are co-authorships of scientific researchers (Globerson et al. 2007), co-occurrences of species (Schmidtlein et al. 2007), connectedness of Enron employees based on their incoming and outgoing email traffic (Klimt and Yang 2004), similarities between shapes resulting from shape context matching (Belongie et al. 2001), similarities of nodes in (scale-free) networks (Breitkreutz et al. 2003), etc.

In this paper, we present a multidimensional scaling technique that attempts to circumvent the two limitations of metric spaces described above when constructing data visualizations, i.e., that is capable of faithfully visualizing non-metric similarity data.[2] Our technique is an extension of a recently proposed technique for multidimensional scaling, called t-Distributed Stochastic Neighbor Embedding (t-SNE; van der Maaten and Hinton 2008). The extension visualizes similarity data by constructing a collection of maps that together represent the similarities between the objects (instead of constructing a single metric map). The new technique, called multiple maps t-SNE, models each object by a point in every map in the collection, and each of the points has an importance weight that indicates its importance in each map. The similarity of two objects under the model is given by a sum over the similarities in all maps, where the similarity in a map depends on both the importance weights of the two points and on their proximity. If two points are close together in a map in which both points have a high importance weight, these points are considered to have a high similarity, even if the points are very far apart in some of the other maps. This "disjunctive" way of working with multiple maps is very different from the standard "conjunctive" approach of using, say, a four-dimensional map and then treating the first two dimensions as one map and the last two dimensions as another map (Roweis and Saul 2000). In the conjunctive approach, a pair of points needs to be close together in *all* of the two-dimensional maps in order to represent high similarity between the corresponding objects. In the disjunctive approach, by contrast, high similarity in one map cannot be vetoed by low similarity in another map.

The visualizations constructed by multiple maps t-SNE can faithfully represent non-metric similarities between objects. For example, the word *tie* can be close to *tuxedo* in a map in which *knot* has a low weight, and close to *knot* in another map in which *tuxedo* a low

[2]We note that our technique should not be confused with traditional techniques for non-metric multidimensional scaling. These techniques aim to rank orders of pairwise distances. See, e.g., Borg and Groenen (2005) for an extensive overview of non-metric multidimensional scaling.

weight. This captures the similarity of *tie* to both *tuxedo* and *knot* without forcing *tuxedo* to be close to *knot*. Moreover, multiple maps t-SNE can (to some extent) model central objects in the data by exploiting the additional space that the multiple maps provide.

An earlier variant of multiple maps t-SNE was proposed in a conference paper by Cook et al. (2007). However, the "aspect maps" model proposed in Cook et al. (2007) did not work very well in practice. The technique presented in this paper works much better in practice, as we show in Sect. 4.3. Moreover, the present paper presents an extensive experimental evaluation of the new technique (the experimental evaluation presented by Cook et al. (2007) is rather limited).

The outline of the remainder of the paper is follows. In Sect. 2, we briefly review multidimensional scaling using t-SNE. Section 3 presents our extension of t-SNE to multiple maps. In Sect. 4, we use multiple maps t-SNE to visualize a word association data set that contains highly non-metric similarity data, and we compare the performance of multiple maps t-SNE with the earlier "aspect maps" model. Section 5 presents experiments in which we use multiple maps t-SNE to visualize machine learning researchers based on their co-authorships of papers published in conference proceedings. Section 7 discusses relations between multiple maps t-SNE and alternative techniques that are designed to deal with non-metric similarities, and it presents directions for future work.

2 t-Distributed stochastic neighbor embedding

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a recently introduced technique for multidimensional scaling (van der Maaten and Hinton 2008) that builds upon earlier work on Stochastic Neighbor Embedding (SNE; Hinton and Roweis 2003; Cook et al. 2007; Globerson et al. 2007). Its input typically consists of a collection of N high-dimensional data vectors $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. In t-SNE, the pairwise distances $\delta_{ij}^2 = \|\mathbf{x}_i - \mathbf{x}_j\|^2$ between the high-dimensional data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are converted into a joint probability distribution P over all pairs of non-identical points. The matrix P has entries

$$p_{ij} = \frac{\exp\left(-\delta_{ij}^2/\sigma\right)}{\sum_k \sum_{l \neq k} \exp\left(-\delta_{kl}^2/\sigma\right)}, \quad \text{for } \forall i \forall j : i \neq j.$$

Since we are only interested in pairwise similarities between points, t-SNE sets $p_{ii} = 0$. Note that for similarity data, such as association or co-occurrence data, the input of t-SNE already naturally takes the form of δ_{ij}'s or p_{ij}'s.

The aim of t-SNE is to model each object by a point $\mathbf{y}_i$ in a low-dimensional map in such a way that the pairwise similarities p_{ij} are modeled as well as possible in the map. We denote the map constructed by t-SNE by $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$. In order to evaluate the pairwise similarities of objects in the map, t-SNE defines joint probabilities q_{ij} that measure the similarity of the points $\mathbf{y}_i$ and $\mathbf{y}_j$ in the low-dimensional map, i.e., q_{ij} is the low-dimensional counterpart of p_{ij}. The error between the input similarities p_{ij} and their counterparts in the low-dimensional map q_{ij} is measured by means of the Kullback-Leibler divergence between the distributions P and Q

$$C(Y) = KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{1}$$

The asymmetric nature of the Kullback-Leibler divergence leads the cost function to focus on appropriately modeling the large pairwise similarities p_{ij} between the input objects. In

other words, similar input objects really need to be close together in the low-dimensional map in order to minimize the cost function $C(Y)$. As the cost function $C(Y)$ is generally non-convex, the minimization of $C(Y)$ is typically performed using a gradient descent method.

The remaining question is how to define the joint probabilities q_{ij} that measure the similarity between the points $\mathbf{y}_i$ and $\mathbf{y}_j$ in the low-dimensional map. A natural choice is to define the q_{ij}'s to be proportional to a Gaussian density, i.e., to define q_{ij} as

$$q_{ij} = \frac{\exp(-\|\mathbf{y}_i - \mathbf{y}_j\|^2)}{\sum_k \sum_{l \neq k} \exp(-\|\mathbf{y}_k - \mathbf{y}_l\|^2)}, \quad \text{for } \forall i \forall j : i \neq j,$$

and to define $q_{ii} = 0$. This definition of the pairwise similarities in the low-dimensional map is used in SNE[3] (Hinton and Roweis 2003), which typically produces fairly good results (see Hinton and Roweis 2003 for some example visualizations). However, SNE suffers from a *crowding problem* that is the result of the exponential volume difference between high and low-dimensional spaces (van der Maaten and Hinton 2008). The crowding problem can be best understood by an example. Suppose that we try to visualize data points that are uniformly sampled from a ten-dimensional hypercube in a two-dimensional map. Also, suppose that our cost function $C(Y)$ is successful in preserving as much of the local structure as possible in the two-dimensional map. Consequently, pairs of points that are only slightly similar have to be modeled too far apart in the map. Since there is a relatively large number of pairs of points that are slightly similar, these points would all like to be closer together in the map. As a result, these pairs of slightly similar points *crush* the low-dimensional map together, which leads to the crowding problem.

The key property of t-SNE is that, in the low-dimensional map, the similarity between two points is not proportional to a Gaussian density, but to that of a Student-t distribution with a single degree of freedom (i.e., a Cauchy distribution)

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}, \quad \text{for } \forall i \forall j : i \neq j,$$

where, again, $q_{ii} = 0$. By using a heavy-tailed distribution to measure similarities in the low-dimensional map, t-SNE allows points that are only slightly similar to be visualized much further apart in the map. This typically leads to very good visualizations (see van der Maaten and Hinton 2008 for example visualizations) compared to alternative techniques for multidimensional scaling. Since its introduction, t-SNE has been successfully applied to the visualization of, among others, documents (Lacoste-Julien et al. 2009), optimization procedures (Erhan et al. 2010), breast cancer CADx data (Jamieson et al. 2010), linguistic data (Mao et al. 2010), paintings (van der Maaten and Postma 2010), and data on malicious software (Gashi et al. 2009; Thonnard et al. 2009). Various extensions and adaptations of t-SNE have been proposed (van der Maaten 2009; Carreira-Perpiñán 2010; Venna et al. 2010; Villmann and Haase 2010; Yang et al. 2010).

[3]Indeed, the original work on SNE uses conditional probabilities $p_{j|i}$ and $q_{j|i}$ instead of joint probabilities p_{ij} and q_{ij}. In practice, using conditional or joint probabilities leads to qualitatively similar results, but the optimization of the joint model requires less computation (van der Maaten and Hinton 2008).

3 Multiple maps t-SNE

The probabilistic nature of t-SNE allows for a natural extension to a multiple maps version, which allows us to circumvent the limitations of metric spaces. Multiple maps t-SNE constructs a collection of M maps, all of which contain N points (one for each of the N input objects). In each map with index m, a point with index i has a so-called importance weight $\pi_i^{(m)}$ that measures the importance of point i in map m. Because of the probabilistic interpretation of our model, we constrain[4] the importance weights $\pi_i^{(m)}$ to make sure that $\forall i \forall m : \pi_i^{(m)} \geq 0$ and $\forall i : \sum_m \pi_i^{(m)} = 1$. We redefine the joint probabilities q_{ij}, which represent the similarities between the objects with index i and j in the visualization, as the weighted sum of the pairwise similarities between the points corresponding to input objects i and j over all M maps. Mathematically, we redefine q_{ij} in the multiple maps t-SNE model as

$$q_{ij} = \frac{\sum_m \pi_i^{(m)} \pi_j^{(m)} (1 + \|\mathbf{y}_i^{(m)} - \mathbf{y}_j^{(m)}\|^2)^{-1}}{\sum_k \sum_{l \neq k} \sum_{m'} \pi_k^{(m')} \pi_l^{(m')} (1 + \|\mathbf{y}_k^{(m')} - \mathbf{y}_l^{(m')}\|^2)^{-1}}, \quad \text{for } \forall i \forall j : i \neq j,$$

where, again, we define $q_{ii} = 0$. The cost function of the multiple maps version of t-SNE is still given by (1), however, it is now optimized with respect to the $N \times M$ low-dimensional map points $\mathbf{y}_i^{(m)}$ and with respect to the $N \times M$ importance weights $\pi_i^{(m)}$.

Because we require the importance weights $\pi_i^{(m)}$ to be positive and we require the importance weights $\pi_i^{(m)}$ for a single point i to sum up to 1 over all maps, direct optimization of the cost function w.r.t. the parameters $\pi_i^{(m)}$ is tedious. To circumvent this problem, we represent the importance weights $\pi_i^{(m)}$ in terms of unconstrained weights $w_i^{(m)}$ (using an idea that is similar to that of softmax units) as follows

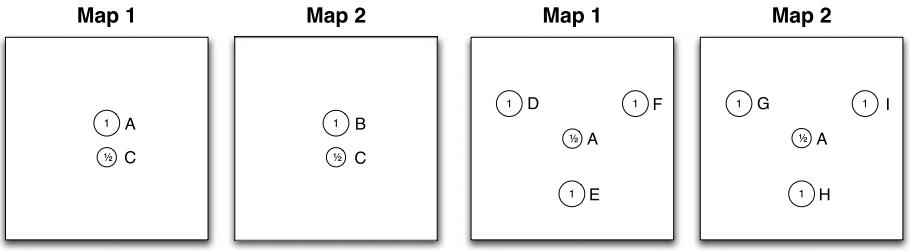
$$\pi_i^{(m)} = \frac{e^{-w_i^{(m)}}}{\sum_{m'} e^{-w_i^{(m')}}}.$$

By defining the importance weights in this way, they are guaranteed to be positive and to sum up to 1. As a result, the minimization of the cost function can be performed with respect to the unconstrained weights $w_i^{(m)}$. This significantly simplifies the optimization of the cost function using gradient descent. The gradients for multiple maps t-SNE are given in Appendix. The details of our gradient descent method are given in Sect. 4.1. Code implementing multiple maps t-SNE is available from http://homepage.tudelft.nl/19j49/multiplemaps.

In contrast to other multidimensional scaling techniques, multiple maps t-SNE can successfully represent (1) intransitive similarities and (2) central objects in two-dimensional visualizations. Below, we explain how multiple maps t-SNE can achieve this.

(1) Intransitive similarities Multiple maps t-SNE can appropriately model intransitive similarities as follows. Assume we have three points A, B, and C that are embedded into two maps (see Fig. 1(a)). Multiple maps t-SNE can give point A an importance weight of 1 in the first map, point B an importance weight of 1 in the second map, and point C an importance weight of $\frac{1}{2}$ in both maps, and it can give all three points nearby spatial locations in both maps. Then, the pairwise similarity between point A and C is roughly equal to

[4]We note here that we might just as well constrain the importance weights $\pi_i^{(m)}$ to sum up to 2 or 12: multiplying the importance weights by a constant scalar value does not change the model in any way.



(a) Intransitive similarities. (b) High centrality.

Fig. 1 Illustration of how multiple maps t-SNE can visualize intransitive similarities and "central" objects

$1 \times \frac{1}{2} = \frac{1}{2}$, and the pairwise similarity between point B and C is also roughly equal to $\frac{1}{2}$. However, the pairwise similarity between point A and B is 0, because the points A and B have no importance weight in each other's maps. Hence, the visualization constructed by multiple maps t-SNE does not satisfy the triangle inequality. As a result, it can visualize intransitive similarities such as those in our introductory example with *tie*, *tuxedo*, and *knot*.

Each time we want to add an object that violates the triangle inequality to an existing map, we need to put one copy of the object in the existing map and another copy in a different map. This way of using multiple maps to violate the triangle inequality is not the same as simply using different maps to visualize different natural clusters in the data. If the triangle inequality is violated by three objects within the same cluster, it is necessary to put two of them in at least two different maps. As a result, visualizations constructed by multiple maps t-SNE typically contain a number of small clusters in each of the maps; clusters within one map that are not adjacent may have little in common.

(2) High centrality Data with high centrality can be visualized appropriately by multiple maps t-SNE, essentially, because multiple maps provide much more space than a single map. We illustrate this by an example. Assume we have six objects that all have the same "central" object A as their most similar object. In a single map, only five of the objects can be modeled such that they have object A as their nearest neighbor. In contrast, when two maps are available, the data can be modeled in such a way that all six objects have object A as their nearest neighbor. For instance, this can be achieved by giving A an importance weight of $\frac{1}{2}$ in both maps, modeling the first three objects close to A in the first map with importance weight 1, and modeling the remaining three objects close to A in the second map with importance weight 1. This example is illustrated in Fig. 1(b). Clearly, the number of points that can have the same point as their nearest neighbor depends on the number of maps and on the dimensionality of the maps.

The reader should note that the multiple maps t-SNE model proposed here is not the same as a *mixture* of t-SNE maps. Indeed, a mixture of maps would employ a single weight per map to measure the importance of that map. Multiple maps t-SNE does not represent the importance of a map,[5] but instead, it represents the importance of each of the words in each of the maps by employing a weight per word per map. Multiple maps t-SNE can thus be viewed as a model in which the similarity representation of each object is modeled by a mixture of similarities in each map (note that because the maps are typically low-dimensional, the similarities within each map are constrained to be low-rank).

[5]Indeed, one could look at the sum of importance weights within a map $\sum_{i=1}^{N} \pi_i^{(m)}$ to obtain a measure of the importance of map m.

4 Visualizing word associations

In this section, we present experiments in which we use multiple maps t-SNE to visualize a large data set of word associations. The setup of these experiments is discussed in Sect. 4.1. Section 4.2 presents the results of the word association experiments. In Sect. 4.3, we compare the performance of multiple maps t-SNE with that of the earlier "aspect maps" (Cook et al. 2007).

4.1 Experimental setup

The word association data set we used in our experiments contains association data for 10,617 words, 5,019 of which were used as input stimuli (Nelson et al. 1998). The data set contains a semantic similarity value for each pair of words that was computed as follows. A large pool of human subjects were given specific words and asked to name associated words. From the subjects' responses, conditional probabilities $p_{j|i}$ are computed that measure the probability that a human subject produces word j in response to word i. The word association data set contains numerous examples of intransitive semantic similarities, such as our introductory example with *tie*, *tuxedo*, and *rope*, and it contains a number of very "central" words that have semantic similarities with many other words in the data (Steyvers and Tenenbaum 2005). As a result, the word association data is a suitable candidate to investigate to what extent multiple maps t-SNE can visualize non-metric similarity data.

In our experiments, we started by symmetrizing and re-normalizing[6] the conditional probabilities $p_{j|i}$ to obtain a joint distribution P that can be used as input into multiple maps t-SNE, i.e., we set $p_{ij} \propto p_{j|i} + p_{i|j}$. Subsequently, we used multiple maps t-SNE to construct 40 maps in which we embed the 5,019 words that were used as input stimuli in the collection of the data (i.e., the 5,019 words i for which we have both the conditional probabilities $p_{j|i}$ and the probabilities $p_{i|j}$). The dimensionality of each map is set to 2.

We trained the model using 2,000 iterations of gradient descent, in which we employed an additional momentum term. In other words, the gradient at each iteration is added to an exponentially decaying sum of the gradients at previous iterations in order to determine the changes in the parameters at each iteration of the gradient search. The momentum term is employed in order to speed up the gradient search without creating the oscillations that are caused by simply increasing the step size. We set the momentum term to 0.5 during the first 250 iterations, and to 0.8 afterwards. For the learning rate, we employed an adaptive learning rate scheme that aims to speed up the optimization by using a different (variable) learning rate for each parameter in the model (Jacobs 1988). In our experiments, we set the initial value of the learning rate to 250 for the map coordinates $\mathbf{y}_i^{(m)}$, and to 100 for the weights $w_i^{(m)}$. Moreover, we employ an approach called "early exaggeration" (van der Maaten and Hinton 2008): in the first 50 iterations of the optimization, we multiply the joint probabilities p_{ij} by 4. As a result, the p_{ij}'s are too large to be appropriately modeled by their corresponding q_{ij}'s (which still sum up to 1). This encourages the optimization to model the largest p_{ij}'s by large q_{ij}'s, thereby creating tight widely separated clusters in the maps that facilitate the identification of an appropriate global organization of the maps. In preliminary experiments, we found the approach to be fairly robust under changes in the optimization parameters. Simpler optimization approaches in which the adaptive learning rate

[6] We could also have used a variant of multiple maps t-SNE that sums over divergences between N conditional distributions, but this has no significant advantages and is computationally more expensive.

and early exaggeration are omitted produce good results as well, but they converge slower. Code implementing our gradient descent optimizer for multiple maps t-SNE is available from http://homepage.tudelft.nl/19j49/multiplemaps.

We visualize the 40 word maps by showing them in an annotated scatter plot, in which the size of a dot represents the importance weight of a word in a specific map. To prevent the visualization from being too cluttered, words with an importance weight below 0.1 were removed from the maps. To increase the legibility of the plots, the annotations in the scatter plots were manually aligned to reduce the overlaps between annotations, while ensuring that word labels are still near their corresponding point in the map.

4.2 Results

Figure 2 shows 6 of the 40 maps that were constructed by multiple maps t-SNE. The full collection of 40 maps can be explored using a web-based interactive visualization tool that implements basic functionalities such as zooming, panning, map search, etc.; this visualization tool is described in more detail in Sect. 6. The reader should note that in the maps in Fig. 2, two words are similar if they are close together in a map in which both words have a high importance weight.

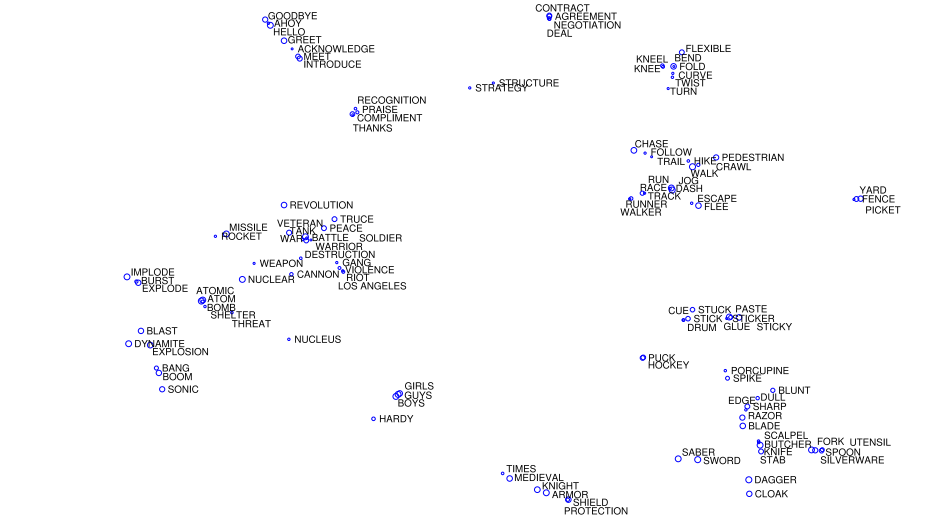
The results presented in Fig. 2 reveal that multiple maps t-SNE retains the similarity structure of the association data fairly well. Because the data contains too many topics, a single map does not generally visualize a single topic. Instead, most maps show two or three main topics, as well as some very small local structures. For instance, map 4 visualizes the topics *sports* and *clothing*, and it shows some small structures that are related to, for instance, the Statue of Liberty: *monument-statue-liberty-freedom*. We note that the maps have a certain "scale" of similarity that depends on the variance of the Student-t distribution in the q_{ij}'s. For instance, map 4 does not indicate that *clothing* is somehow related to the Statue of Liberty, because the two topics are widely separated in the map.

The results reveal how multiple maps t-SNE circumvents[7] the limitations of metric spaces when constructing a visualization of non-metric similarity data. In particular, the maps reveal many of the intransitive similarities of words. For instance, the semantic similarity of the word *tie* to words such as *rope* and *knot* is modeled in map 2, whereas in map 4, the semantic similarity of the word *tie* with *suit*, *tuxedo*, and *prom* is modeled. In addition, map 5 reveals the semantic similarity of *tie* with words such as *ribbon* and *bow*. As a second example, the semantic similarity of the word *cheerleader* with various kinds of sports is modeled in map 4, whereas map 6 reveals the association of *cheerleader* with words such as *gorgeous*, *beauty*, and *sexy*. A third example is the word *monarchy*, which is shown close to words that are related to royalty such as *king*, *queen*, *crown*, and *royal* in map 3. In map 6, the word *monarchy* is close to other governmental forms such as *oligarchy*, *anarchy*, *democracy*, and *republic*.

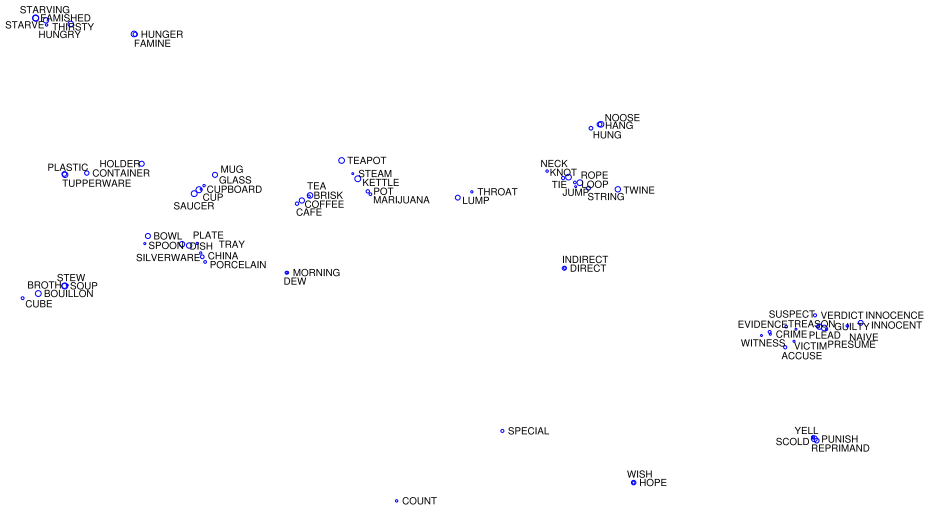
4.3 Comparison with aspect maps

In earlier work, Cook et al. (2007) proposed a technique called "aspect maps" that is very similar to multiple maps t-SNE, however, it (1) uses Gaussian instead of Student-t densities in the definition of q_{ij} and (2) uses asymmetric similarities $q_{j|i}$ instead of our symmetric

[7]We should note that it is hard to find a good example of the visualization of "central" words without closely studying all 40 maps. Hence, we focus on intransitive similarities in our discussion of the experimental results.



(a) Map 1.

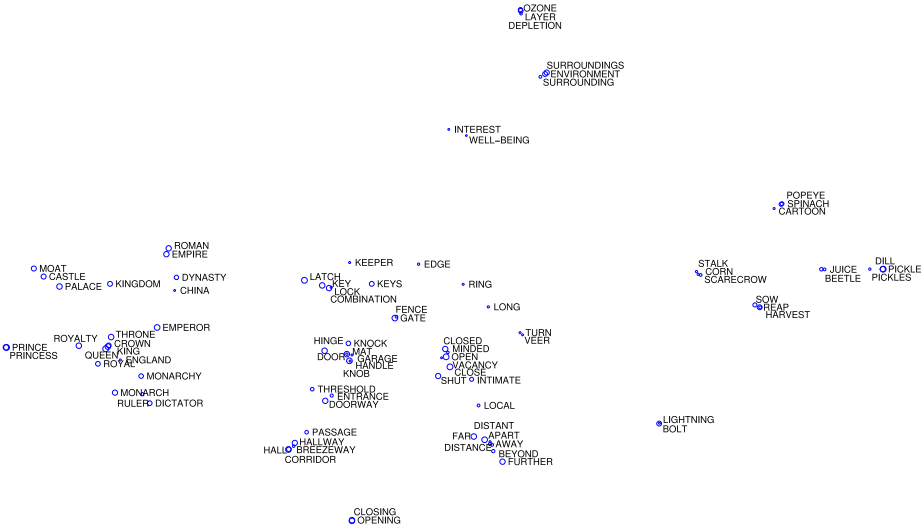


(b) Map 2.

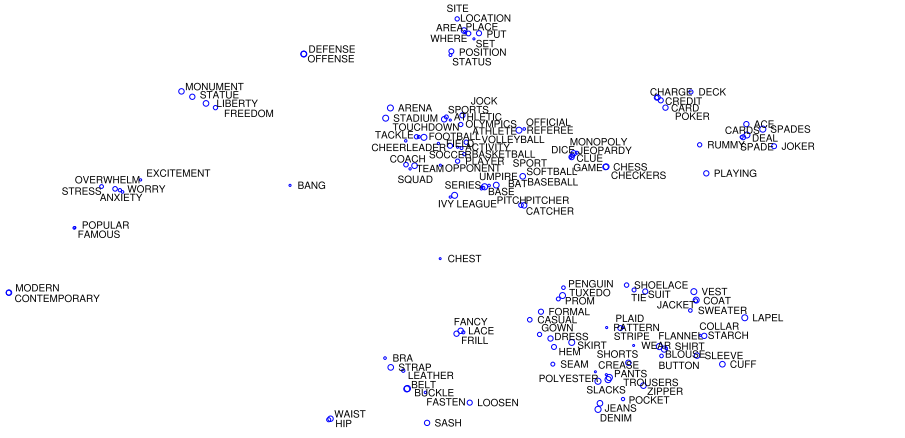
Fig. 2 Results of multiple maps t-SNE on the word association data set (**a**–**b**). Because of space limitations, we only show 6 of the original 40 maps

similarities q_{ij}. We compare[8] the performance of aspect maps and multiple maps t-SNE in experiments on the 1,000 most frequent words in the association data. In our comparative

[8]We also performed preliminary experiments with an approach in which we use t-SNE to construct a representation of, say, 20 dimensions, and use that representation to make 10 two-dimensional maps. Such an approach has the problem that words that have no semantic similarity whatsoever can be close together in a map (as long as these words are apart in some of the other dimensions, the original 20-dimensional represen-



(c) Map 3.

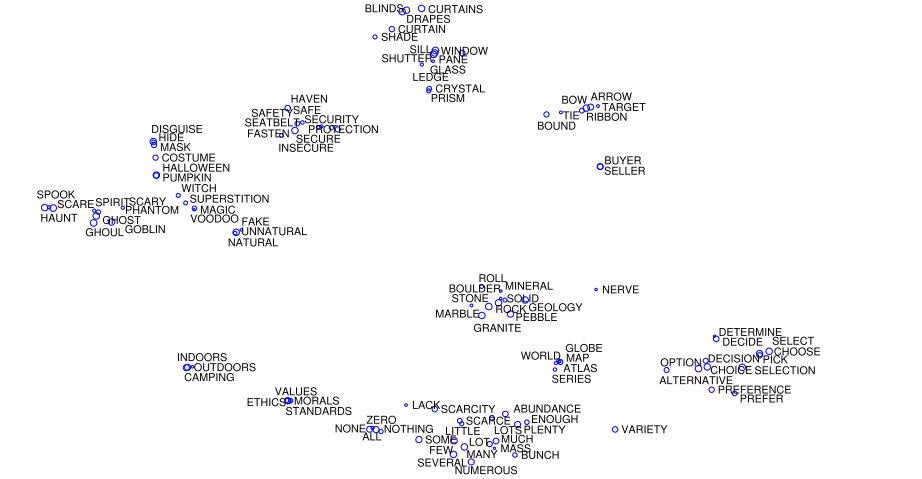


(d) Map 4.

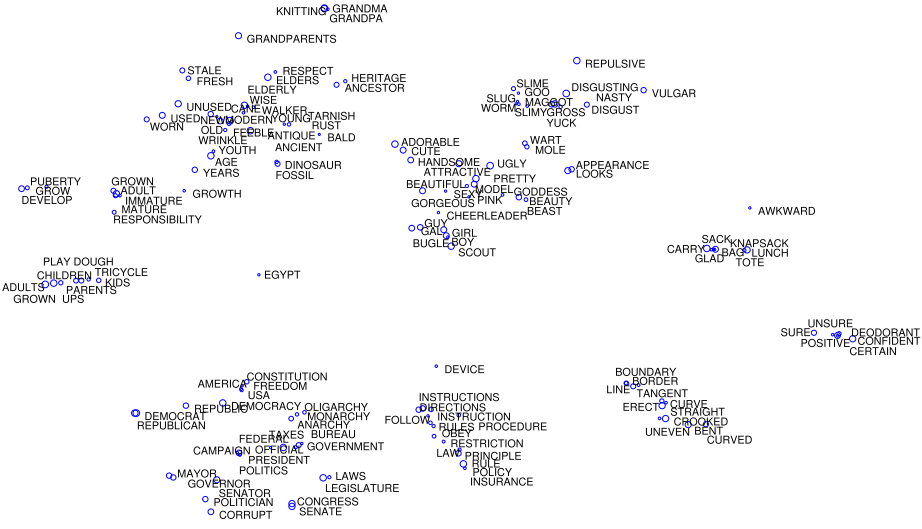
Fig. 2 Results of multiple maps t-SNE on the word association data set (**c**–**d**). Because of space limitations, we only show 6 of the original 40 maps

experiments, we used exactly the same optimizer for both aspect maps and multiple maps t-SNE (i.e., we used a gradient descent algorithm with momentum and early exaggeration). To assess how well the input similarities are modeled by a multiple maps model, we measure its *neighborhood preservation ratio*: for each word i, we measure the ratio of the k most similar words in the association data that are modeled as nearest neighbors under the

tation models the underlying structure correctly). This leads to maps that show completely arbitrary structure, which is why do not show the results of such an approach here.



(e) Map 5.



(f) Map 6.

Fig. 2 Results of multiple maps t-SNE on the word association data set (**e**–**f**). Because of space limitations, we only show 6 of the original 40 maps

multiple maps model. In other words, for word i, we measure the ratio of words j with the k highest q_{ij}-values that are among the words j with the k highest p_{ij}-values. We average the neighborhood preservation ratio over all words i. Although it is unlikely that an observer would go through a calculation similar to the computation of the q_{ij}-values, the neighborhood preservation ratio does in some sense correspond to the way in which users may use multiple maps models. Specifically, a typical observer will go through the maps to identify the different types of similarities/relations that an object has. In other words, an observer

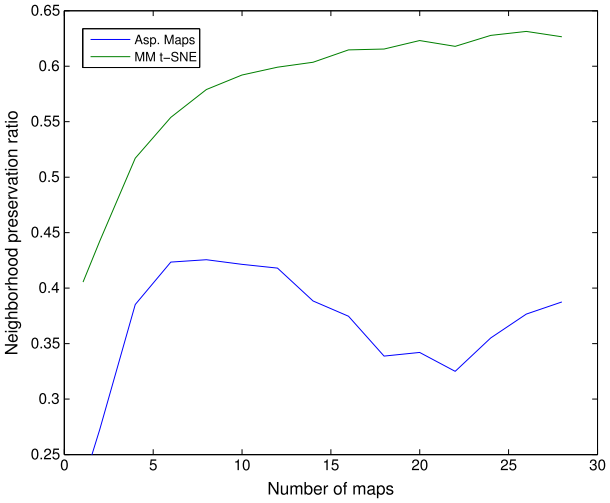


Fig. 3 Neighborhood preservation ratio for aspect maps and multiple maps t-SNE on the word association data as a function of the number of two-dimensional maps

tends to collect all similarities of an object;[9] the neighborhood preservation ratio measures to what extent these similarities are correct. Indeed, a measure similar to the neighborhood preservation ratio has previously been proposed by Venna et al. (2010) as an appropriate measure of how observers perceive single-map visualizations.

In Fig. 3, the neighborhood preservation ratio (measured using $k = 1$) for aspect maps and multiple maps t-SNE is plotted as a function of the number of maps. The results reveal that multiple maps t-SNE outperforms aspect maps by a large margin. Multiple maps t-SNE gets more than half of all neighborhood relations right using just three two-dimensional maps, whereas none of the aspect maps solutions achieves an neighborhood preservation ratio of 0.5 or higher. Moreover, the results in Fig. 3 reveal that multiple maps t-SNE is capable of exploiting the additional space that becomes available when extra maps are added: adding maps facilitates better preservation of neighborhood relations under the model. We also measured neighborhood preservation ratios for $k = 3, 5, \ldots, 17$: the results of these experiments were very similar, which is why we omitted them here.

The strong performance of multiple maps t-SNE in terms of neighborhood preservation ratio compared to aspect maps can be largely explained by its ability to construct large separations between different topics within the same map. As a result, multiple topics can be visualized in a single map without distorting the neighborhood relations. Moreover, the use of a heavy-tailed distribution to measure similarities in the maps tends to make gradient descent easier (van der Maaten and Hinton 2008).

An interesting aspect of the neighborhood preservation ratio that is illustrated by the graph in Fig. 3 is that it allows us to make an informed choice of the number of maps to be used in the multiple maps visualization. The graph suggests that the neighborhood preservation ratio has an asymptotic behavior: after the model has a particular number of maps, adding new maps does not appear to lead to improved performance, suggesting that (1) the new maps are representing similarity structure that was already modeled in some of the other maps and/or (2) the new maps simply 'split up' old maps into two subsets of objects

[9]To browse through all maps efficiently, the observer may employ interactive visualization tools. We discuss such tools in more detail in Sect. 6.

that are mostly unrelated. For the word association data, a collection of approximately 15 maps appears to provide sufficient space to model the main similarity structure in the data.

5 Visualizing NIPS co-authorships

In this section, we present the results of experiments in which we use multiple maps t-SNE to visualize machine learning researchers based on their co-authorships of papers published in the proceedings of the annual Neural Information Processing Systems (NIPS) conference. The setup of the experiments is presented in Sect. 5.1. Section 5.2 presents the results of the experiments.

5.1 Experimental setup

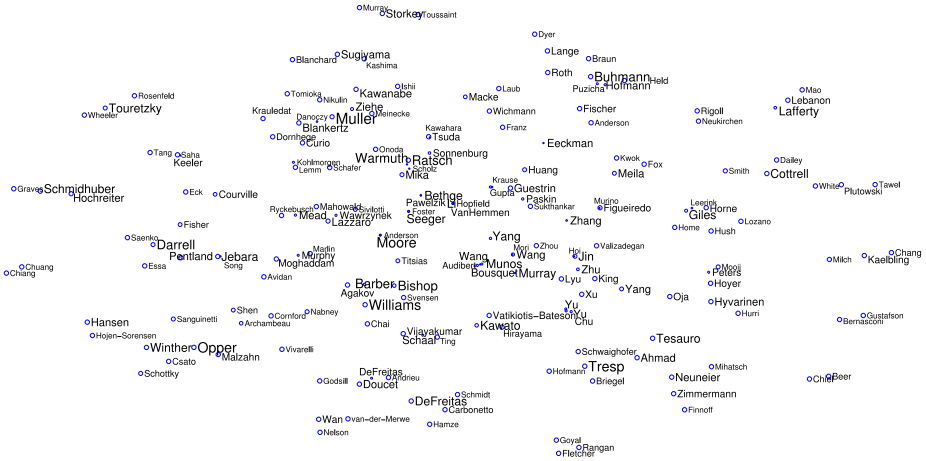
We collected data on the authors of all papers that appeared in NIPS volume 1 to 22 (i.e., between 1988 and 2009) from the NIPS website. Because we are interested in visualizing co-authorships of machine learning researchers, we preprocessed the data by removing all single-author papers. We also eliminated relatively "unimportant" researchers from the data by removing authors that only have a single paper in NIPS. After these preprocessing steps,[10] the data set contains 1,418 authors who together wrote 2,121 papers in NIPS. Each of the authors has an average of 4.3 papers in the NIPS proceedings, and each of the 2,121 papers has an average of 2.5 authors. The data set is available for download from http://homepage.tudelft.nl/19j49/multiplemaps. The data set contains a large number of intransitive similarities, because researcher A may have co-authored one or more papers with researcher B, and researcher B may have co-authored papers with researcher C, whilst researcher A and C have never collaborated on a paper. This happens, in particular, when researchers accept a position at a different institution during their career, or when researchers change their research interests. The large number of intransitive similarities in the NIPS co-authorship data set makes it a suitable candidate to investigate the performance of multiple maps t-SNE.

We computed a co-authorship matrix from the preprocessed data and normalized it to obtain conditional probabilities $p_{j|i}$, which indicate the conditional probability that, given that author i is an author of a NIPS paper, author j is also an author of that same NIPS paper. As in the experiments with the word association data, we symmetrized the conditional probabilities (i.e., we set $p_{ij} \propto p_{j|i} + p_{i|j}$), and we used the resulting joint probabilities as input into multiple maps t-SNE. We used multiple maps t-SNE to construct 10 two-dimensional maps. The parameters of the optimizer (i.e., learning rate, momentum, etc.) were set to exactly the same values as in the word association experiments.

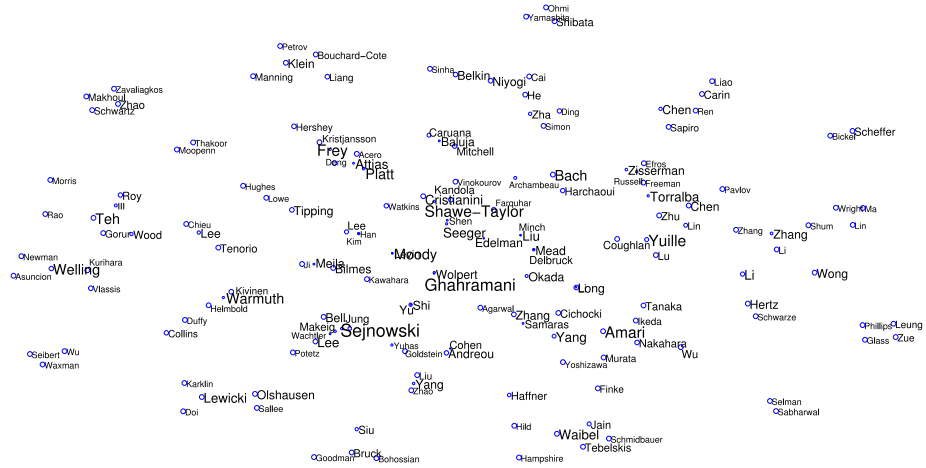
5.2 Results

We present 4 of the 10 resulting maps in Fig. 4. The font size of an author's name in the maps is proportional to the logarithm of his/her total number of papers in NIPS (i.e. the font size indicates the *importance* of an author in the NIPS community), whereas the size of a

[10] We should note here that the parsing of author names is problematic, for instance, because there are some authors who have the same last name and their first name starts with the same character too, or because the spelling of authors' first names may vary.



(a) Map 1.

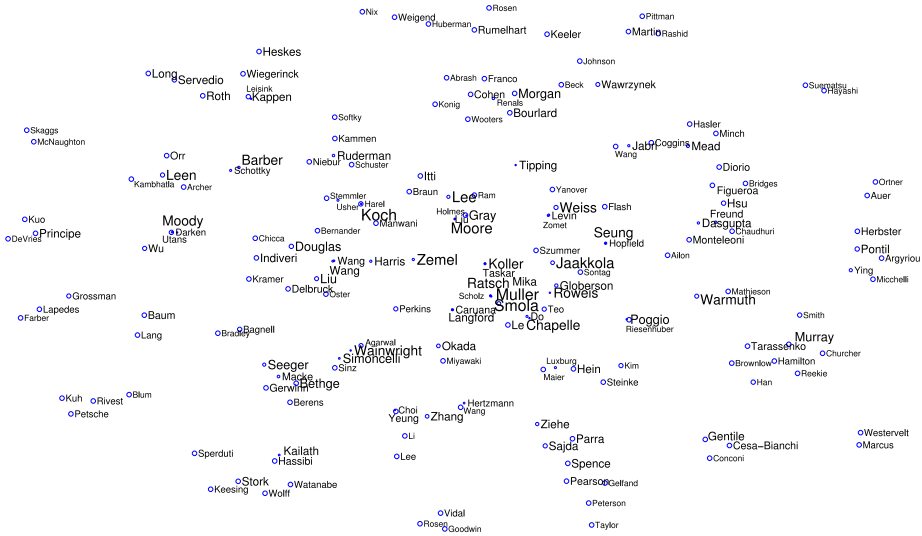


(b) Map 2.

Fig. 4 Results of multiple maps t-SNE on the NIPS co-authorship data set (map 1 and 2)

dot indicates the importance of the author in the map. The full collection of 10 maps can be explored using the online visualization tool described in Sect. 6.

Similar to the results on the word association data presented in the previous section, each of the maps constructed from the NIPS co-authorship data does not visualize a separate cluster, i.e., there are no separate maps for, e.g., "neural networks researchers", "Bayesians", or "manifold learners". Instead, each map shows a few "cliques" of researchers who intensively cooperate, as well as smaller structures of researchers who wrote only one or two papers together in NIPS. The NIPS author maps do successfully capture (part of) the non-metric structure in the data. Importantly, the maps reveal researchers whose collaborators changed over time, for instance, researchers who worked for quite some time in a specific research lab and then moved to another research lab. For example, in map 4, Max Welling is shown close to his collaborators in Toronto (where he did his post-doc), whereas he is shown close



(c) Map 3.

(d) Map 4.

Fig. 4 Results of multiple maps t-SNE on the NIPS co-authorship data set (map 3 and 4)

to collaborators from UC Irvine (where he is currently a professor) in map 2. As a second example, Martin Wainwright has collaborated extensively with both Eero Simoncelli and Michael Jordan, but on different topics and at different times. He appears with Simoncelli in map 3 and with Jordan in map 4 thus allowing their representations to remain far apart. As a third example, Klaus-Robert Müller's collaborations until 2000 (with, among others, Alex Smola and Gunnar Rätsch) are visualized in map 3, whereas his collaborations after 2000 (for instance, with Benjamin Blankertz) are shown in map 1.

Figure 5 shows the neighborhood preservation ratio obtained by aspect maps and multiple maps t-SNE for increasing numbers of maps. The results presented in the figure are in

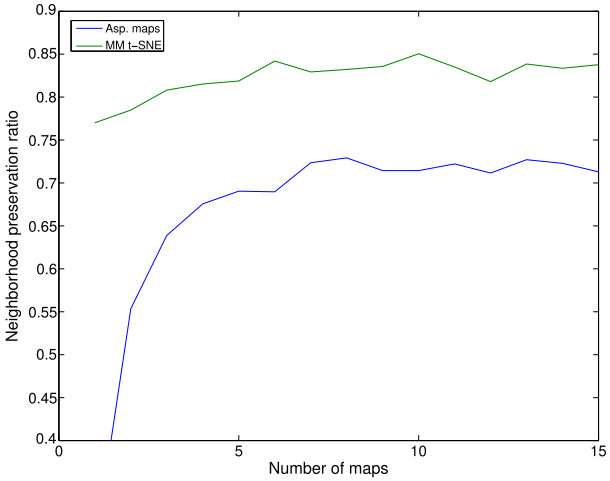


Fig. 5 Neighborhood preservation ratio for aspect maps and multiple maps t-SNE on the NIPS co-authorship data as a function of the number of two-dimensional maps

line with those presented in Sect. 4.3: multiple maps t-SNE consistently outperforms aspect maps. The results also suggest that appropriately visualizing NIPS authors require significantly less maps than visualizing word association; approximately 5 maps appear to suffice for modeling the co-authorship. Presumably, visualizing co-authorships requires fewer maps than visualizing word associations because there are far fewer authors than words in the respective data sets.

6 Interpretability

An important issue that arises when constructing a multiple maps t-SNE model with, say, 40 maps is the interpretability of such models. In particular, rapid navigation and interpretation of the maps becomes increasingly problematic as the number of maps increases. To address this problem, we envision a combination of multiple maps t-SNE with state-of-the-art techniques for information visualization and visual analytics (Thomas and Cook 2005; Keim et al. 2010). To illustrate the potential of such a combination, we developed a web-based visualization that allows users to rapidly navigate large numbers of maps. The visualization was developed using the D3 framework (Bostock et al. 2011); D3 is an information visualization framework that facilitates binding arbitrary data (such as the multiple maps) to a Document Object Model (DOM) and subsequently applying data-driven transformations on the resulting document. The web-based visualizations of both the word associations and of the NIPS authors are available online at http://homepage.tudelft.nl/19j49/multiplemaps.

The most typical way in which an observer employs multiple maps models is a by querying for all occurrences of an object (such as a word or a NIPS author) in the collection of maps, in order to discover the different types of relations of that object that are modeled in the different maps. Interactive visualizations such as the one we developed facilitate such queries through simple interactions. For instance, an observer may type the word in a search box (or otherwise specify the object of interest) to display all maps in which that word occurs (with sufficiently high importance weight) in a single screen; the maps are automatically panned and zoomed in order to reveal the location of the word in that map. Another interaction allows the observer to click on an object in the currently displayed map(s) in order to obtain information about in which maps that object is modeled with a sufficiently high importance weight; a second click shows all these maps in a single screen.

The web-based visualization we developed is merely meant as an illustration of how navigating multiple maps may become practical. Indeed, the visualization may be extended and improved in various ways. For instance, right-clicking on an object could reveal relevant information about that object (such as the titles of all published NIPS papers by an author), parts of the word association or co-authorship graph could be overlaid on the visualizations, etc. When deploying a multiple maps visualization tool, the usability of that tool should be investigated through user studies, as is common practice in the information visualization community (Plaisant 2004).

7 Discussion

In Sects. 4 and 5, we presented the results of experiments that reveal that multiple maps t-SNE can successfully visualize non-metric similarities, and in particular, intransitive similarities. In this section, we discuss the relations of multiple maps t-SNE with other techniques that can deal with non-metric data. Specifically, we discuss the similarities and differences of multiple maps t-SNE with (1) techniques that consider the negative part of the eigenspectrum of the Gram matrix and (2) topic models such as Latent Dirichlet Allocation (LDA; Blei et al. 2003).

Multiple maps t-SNE has a similar goal to techniques that exploit the structure from the eigenvectors that correspond to the negative eigenvalues of a (centered) pairwise distance matrix (the so-called Gram matrix). The eigenvectors corresponding to negative eigenvalues contain structural information on the metric violations in the pairwise dissimilarity matrix (Laub and Müller 2004; Laub et al. 2007). However, an approach that exploits the negative eigenspectrum has two main disadvantages compared to multiple maps t-SNE. First, in contrast to multiple maps t-SNE, approaches that employ the negative part of the eigenspectrum can only construct two metric maps:[11] a "positive map" and a "negative map". Second, it is hard to interpret the map that corresponds to the negative part of the eigenspectrum: the map that corresponds to the positive part of the eigenspectrum is a metric approximation to the similarities, and the negative map is constructed in such a way as to correct the errors in the positive map. As a result, the negative map generally also contains a lot of noise.

Multiple maps t-SNE has interesting connections to topic models such as Latent Dirichlet Allocation (LDA; Blei et al. 2003) and others (e.g., McCallum 1999; McCallum et al. 2004; Rosen-Zvi et al. 2004). LDA is a generative model in which each word x is drawn from a topic z which is in turn picked from a multinomial distribution over k topics. This distribution over topics θ is in turn drawn from a Dirichlet distribution. The key characteristic of LDA is that the k topics are multinomial distributions over words.[12] The topics can be used for visualization of data, for instance, by listing the most probable words under each topic in a table (see, e.g., Blei et al. 2003).

Under a topic model, two objects can be viewed as similar if they both have a high probability under at least one of the k topics (Griffiths et al. 2007). This provides topic models with the same desirable properties that multiple maps t-SNE has (although topic models cannot

[11] We note it is possible to consider these two maps as a single hyperbolic space in which distance measures can be defined that do not obey the triangle inequality (Pekalska and Duin 2005), however, this does not lead to intuitive visualizations.

[12] The variable k is a parameter that sets the number of topics that is employed in the semantic representation. It may either be set by the user, or it may be learned from the data using non-parametric Bayesian techniques (Blei et al. 2004; Teh et al. 2004).

lay out the input objects in a map). In particular, a topic model is capable of modeling intransitive semantic similarities in different topics. Analogous to our example with *tie*, *tuxedo*, and *knot*, in topic models, *tie* and *tuxedo* could be given a high probability in one topic and *tie* and *knot* could be given high probability in another topic, which would not make *tuxedo* similar to *knot* under the model. In the same way, topic models can model central objects by giving them a high probability in a large number of topics, which automatically gives rise to asymmetric similarities. Indeed, the advantages of multiple maps t-SNE over traditional multidimensional scaling techniques are comparable to the advantages of topic models over semantic space models such as Latent Semantic Analysis (LSA; Landauer and Dumais 1997; Hofmann 1999). The advantages of topic models over LSA are described in detail by, e.g., Griffiths et al. (2007).

The main difference between topic models and multiple map t-SNE is that, in contrast to topic models, multiple map t-SNE can (1) be trained directly on association or co-occurrence data and (2) capture subtle semantic structure in the spatial structure of the maps. The first capability may be relevant depending on the input data that is available.[13] The merits of the second capability are illustrated, for instance, in the 'sports' cluster in map 4 of Fig. 2, where the subtle semantic difference between physical sports such as *football*, *baseball*, and *volleybal*, and mental sports such as *chess*, *checkers*, and *poker* is captured in the spatial structure of the cluster (from left to right). Topic models cannot faithfully visualize such subtle differences because they only have two options: grouping the words together in a topic or not. In addition, multiple map SNE has the advantage that it can model small semantic structures that are not closely related to other semantic structures, such as the *Popeye-spinach-cartoon* cluster in Fig. 2(c), without resorting to the construction of a new map or topic.

A minor disadvantage of multiple maps t-SNE is that it is not tailored to clustering the concepts in the data, i.e., concepts that have a high importance weight in the same map do not necessarily all correspond to the same cluster or topic. Multiple maps t-SNE can thus not be viewed as combining techniques for finding overlapping clusters (Banerjee et al. 2005; Heller and Ghahramani 2007) with techniques for embedding. The reason for this lies in the structure of the cost function in (1): due to the asymmetry of the Kullback-Leibler divergence, the cost function does not severely penalize cases in which dissimilar objects (low p_{ij}) both have a high importance weight in the same map (high q_{ij}). We found in preliminary experiments that, in cases in which it is desirable that maps only model a single topic, it is better to minimize the inverse Kullback-Leibler divergence $KL(Q||P)$ instead of the "normal" divergence $KL(P||Q)$. However, one should note that this may have a negative influence on the spatial layout of the maps, since in terms of spatial layout of the maps, the inverse Kullback-Leibler divergence will focus on modeling dissimilar objects far apart (i.e., focus on global similarity structure) instead of on modeling similar objects close together (i.e., focus on local similarity structure).

A remaining problem is how to select the number of maps to use in a multiple maps model. A similar problem occurs in clustering, where the number of clusters needs to be set. A simple approach to address the model selection problem is by monitoring the effect of adding maps on the neighborhood preservation ratio, and selecting the number of maps at which the neighborhood preservation ratio starts to decay. An alternative approach to the model selection problem, that recently has become popular in clustering, is to monitor the stability of solutions (von Luxburg 2010). Such an approach

[13] We note that word similarities can be learned in an unsupervised way from text corpora, see, e.g., Lund et al. (1995), Collobert and Weston (2008), Mnih and Hinton (2009).

chooses the number of maps in a such way that the corresponding results are most stable under small perturbations of the input similarities. Indeed, an approach that automatically adds extra maps during the optimization if this is required to model the data, like in affinity propagation (Frey and Dueck 2007) or non-parametric Bayesian techniques (Blei et al. 2004; Teh et al. 2004), may be more appealing. We have experimented with such approaches, but hitherto, we have not found them to produce good results.

A possible direction for future work on multiple maps t-SNE is to investigate alternative combinations of clustering and visualization, which introduce a penalty term that penalizes "impurity" of the maps. For instance, one could force all points that have high importance weight in the same map to be slightly similar by introducing a small background similarity λ in the maps by setting

$$q_{ij} \propto \sum_m \left[\pi_i^{(m)} \pi_j^{(m)} \left(\lambda + (1-\lambda)\left(1 + \|\mathbf{y}_i^{(m)} - \mathbf{y}_j^{(m)}\|^2\right)^{-1}\right)\right]$$

with $0 \leq \lambda \leq 1$. Other interesting directions for future work include variants of multiple maps t-SNE in which the importance weights are not required to sum up to 1, but are only required to be positive and smaller than 1. Using such a definition of the importance weights, the q_{ij}'s are still probabilities. Dropping the requirement that the importance weights sum up to 1 may lead to visualizations that are better at modeling "central" objects in the data, because the central objects can be given a high importance weight in many of the maps. As we mainly use multiple maps to identify different aspects of the local structure around an object, future work may also focus on developing objective functions or constraint sets that explicitly aim at identifying such aspects of local structure (and that do not construct global maps that contain all input objects).

Acknowledgements Parts of this work was performed while Laurens van der Maaten was affiliated to University of California, San Diego and to Tilburg University, The Netherlands. Laurens van der Maaten is supported by the Netherlands Organization for Scientific Research (NWO; Rubicon grant No. 680.50.0908) and by the EU-FP7 Network of Excellence on Social Signal Processing (SSPNet). Geoffrey Hinton is a fellow of the Canadian Institute for Advanced Research, and is also supported by grants from NSERC and CFI and gifts from Google and Microsoft.

Appendix: Gradients of multiple maps t-SNE

Recall that multiple maps t-SNE tries to minimize the cost function

$$C(Y) = KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}},$$

where the pairwise similarities under the model q_{ij} are defined as

$$\forall i \forall j, \quad i \neq j : q_{ij} = \frac{\sum_m \pi_i^{(m)} \pi_j^{(m)} (1 + \|\mathbf{y}_i^{(m)} - \mathbf{y}_j^{(m)}\|^2)^{-1}}{\sum_k \sum_{l \neq k} \sum_{m'} \pi_k^{(m')} \pi_l^{(m')} (1 + \|\mathbf{y}_k^{(m')} - \mathbf{y}_l^{(m')}\|^2)^{-1}},$$

where the importance weights $\pi_i^{(m)}$ are expressed in terms of unconstrained weights $w_i^{(m)}$

$$\pi_i^{(m)} = \frac{e^{w_i^{(m)}}}{\sum_{m'} e^{w_i^{(m')}}}.$$

To simplify the notation of the gradients, we define

$$d_{ij}^{(m)} = \|\mathbf{y}_i^{(m)} - \mathbf{y}_j^{(m)}\|^2,$$
$$Z = \sum_k \sum_{l \neq k} \sum_{m'} \pi_i^{(m')} \pi_k^{(m')} \big(1 + d_{kl}^{(m')}\big)^{-1}.$$

The gradient of the cost function with respect to the low-dimensional map point $\mathbf{y}_i^{(m)}$ is given by

$$\frac{\partial C(Y)}{\partial \mathbf{y}_i^{(m)}} = 4 \sum_j \frac{\partial C(Y)}{\partial d_{ij}^{(m)}} \big(\mathbf{y}_i^{(m)} - \mathbf{y}_j^{(m)}\big),$$

where the gradient with respect to the squared Euclidean distance between $\mathbf{y}_i^{(m)}$ and $\mathbf{y}_j^{(m)}$ in map m is given by

$$\frac{\partial C(Y)}{\partial d_{ij}^{(m)}} = \frac{\pi_i^{(m)} \pi_j^{(m)} (1 + d_{ij}^{(m)})^{-1}}{q_{ij} Z} (p_{ij} - q_{ij}) \big(1 + d_{ij}^{(m)}\big)^{-1}.$$

The gradient of the cost function with respect to the importance weights $\pi_i^{(m)}$ is given by

$$\frac{\partial C(Y)}{\partial w_i^{(m)}} = \pi_i^{(m)} \left(\left(\sum_{m'} \pi_i^{(m')} \frac{\partial C(Y)}{\partial \pi_i^{(m')}} \right) - \frac{\partial C(Y)}{\partial \pi_i^{(m)}} \right),$$

where the gradient of the cost function with respect to the importance weights $\pi_i^{(m)}$ is given by

$$\frac{\partial C(Y)}{\partial \pi_i^{(m)}} = \sum_j \left(\frac{2}{q_{ij} Z} (p_{ij} - q_{ij}) \right) \pi_j^{(m)} \big(1 + d_{ij}^{(m)}\big)^{-1}.$$

References

Banerjee, A., Krumpelman, C., Basu, S., Mooney, R., & Ghosh, J. (2005). Model based overlapping clustering. In *Proceedings of the ACM SIGKDD conference on knowledge discovery and data mining*.

Belkin, M., & Niyogi, P. (2002). Laplacian Eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems* (Vol. 14, pp. 585–591).

Belongie, S., Malik, J., & Puzicha, J. (2001). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(4), 509–522.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. In S. Thrun, L. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems* (Vol. 16, pp. 17–24). Cambridge: The MIT Press.

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling* (2nd ed.). New York: Springer.

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D3: Data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2301–2309.

Breitkreutz, B.-J., Stark, C., & Tyers, M. (2003). Osprey: a network visualization system. *Genome Biology*, *4*(3), R22.1–R22.4.

Carreira-Perpiñán, M. Á. (2010). The elastic embedding algorithm for dimensionality reduction. In *Proceedings of the* 27*th international conference on machine learning* (pp. 167–174).

Cayton, L., & Dasgupta, S. (2006). Robust Euclidean embedding. In *Proceedings of the* 23*rd international conference on machine learning* (pp. 169–176).

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the international conference on machine learning* (pp. 160–167).

Cook, J. A., Sutskever, I., Mnih, A., & Hinton, G. E. (2007). Visualizing similarity data with a mixture of maps. *JMLR Workshop and Conference Proceedings*, *2*, 67–74.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., & Bengio, S. (2010) Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, *11*, 625–660.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*, 972–976.

Gashi, I., Stankovic, V., Leita, C., & Thonnard, O. (2009). An experimental study of diversity with off-the-shelf antivirus engines. In *Proceedings of the IEEE international symposium on network computing and applications* (pp. 4–11).

Globerson, A., & Roweis, S. (2007). Visualizing pairwise similarity via semidefinite programming. In *Proceedings of the* 11*th international workshop on artificial intelligence and statistics (AI-STATS)* (pp. 139–146).

Globerson, A., Chechik, G., Pereira, F., & Tishby, N. (2007). Euclidean embedding of co-occurrence data. *Journal of Machine Learning Research*, *8*, 2265–2295.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. L. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

Heller, K. A., & Ghahramani, Z. (2007). A nonparametric Bayesian approach to modeling overlapping clusters. In *Proceedings of the* 11*th international conference on artificial intelligence and statistics*.

Hinton, G. E., & Roweis, S. T. (2003). Stochastic neighbor embedding. In *Advances in neural information processing systems* (Vol. 15, pp. 833–840).

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the* 22*th annual international SIGIR conference* (pp. 50–57). New York: ACM Press.

Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Networks*, *1*, 295–307.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008). Similarity, kernels, and the triangle inequality. *Journal of Mathematical Psychology*, *52*(2), 297–303.

Jamieson, A. R., Giger, M. L., Drukker, K., Li, H., Yuan, Y., & Bhooshan, N. (2010). Exploring nonlinear feature space dimension reduction and data representation in breast CADx with Laplacian Eigenmaps and t-SNE. *Medical Physics*, *37*(1), 339–351.

Keim, D. A., Kohlhammer, J., Ellis, G., & Mansmann, F. (2010). *Mastering the information age; solving problems with visual analytics*. Eurographics Association.

Klimt, B., & Yang, Y. (2004). *Lecture notes in computer science: Vol. 3201. The Enron corpus: a new dataset for email classification research* (pp. 217–226).

Kruskal, J. B., & Wish, M. (1986). *Multidimensional scaling*. Beverly Hills: Sage.

Lacoste-Julien, S., Sha, F., & Jordan, M. I. (2009). DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Advances in neural information processing systems* (Vol. 21, pp. 897–904).

Lafon, S., & Lee, A. B. (2006). Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(9), 1393–1403.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Laub, J., & Müller, K.-R. (2004). Feature discovery in non-metric pairwise data. *Journal of Machine Learning Research*, *5*, 801–818.

Laub, J., Macke, J., Müller, K.-R., & Wichmann, F. A. (2007). Inducing metric violations in human similarity judgements. In *Advances in neural information processing systems* (Vol. 19, pp. 777–784).

Lawrence, N. D. (2011). Spectral dimensionality reduction via maximum entropy. In *Proceedings of the international conference on artificial intelligence and statistics* (pp. 51–59).

Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the* 17*th annual conference of the cognitive science society* (pp. 660–665). Mahwah: Erlbaum.

Mao, Y., Balasubramanian, K., & Lebanon, G. (2010). Dimensionality reduction for text using domain knowledge. In *Proceedings of the* 23*rd international conference on computational linguistics* (pp. 801–809).

McCallum, A. (1999). Multi-label text classification with a mixture model trained by em. In *AAAI workshop on text learning*. New York: ACM.

McCallum, A., Corrada-Emmanuel, A., & Wang, X. (2004). *The author-recipient-topic model for topic and role discovery in social networks: experiments with Enron and academic email* (Technical Report UM-CS-2004-096). Department of Computer Science, University of Massachusetts, Amherst, MA.

Mnih, A., & Hinton, G. E. (2009). A scalable hierarchical distributed language model. In *Advances in neural information processing systems* (pp. 1081–1088).

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*.

Pekalska, E., & Duin, R. P. W. (2005). *The dissimilarity representation for pattern recognition: foundations and applications*. Singapore: World Scientific.

Plaisant, C. (2004). The challenge of information visualization evaluation. In *Proceedings of the working conference on advanced visual interfaces*.

Rosen-Zvi, M., Griffiths, T., Steyversand, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the* 20*th conference on uncertainty in artificial intelligence*. Arlington: AUAI Press.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by Locally Linear Embedding. *Science*, *290*(5500), 2323–2326.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, *18*(5), 401–409.

Schmidtlein, S., Zimmermann, P., Schüpferling, R., & Weiss, C. (2007). Mapping the floristic continuum: ordination space position estimated from imaging spectroscopy. *Journal of Vegetation Science*, *18*, 131–140.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge: MIT Press.

Schölkopf, B., Smola, A. J., & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, *10*(5), 1299–1319.

Shaw, B., & Jebara, T. (2009). Structure preserving embedding. In *Proceedings of the international conference on machine learning* (pp. 937–944).

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41–78.

Teh, Y., Jordan, M. I., Beal, M., & Blei, D. M. (2004). Hierarchical Dirichlet processes. In *Advances in neural information processing systems* (Vol. 17, pp. 1385–1392). Cambridge: MIT Press.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323.

Thomas, J. J., & Cook, K. A. (2005). *Illuminating the path: the research and development agenda for visual analytics*.

Thonnard, O., Mees, W., & Dacier, M. (2009). Addressing the attack attribution problem using knowledge discovery and multi-criteria fuzzy decision-making. In *Proceedings of the ACM SIGKDD workshop on CyberSecurity and intelligence informatics* (pp. 11–21).

Torgerson, W. S. (1952). Multidimensional scaling I: theory and method. *Psychometrika*, *17*, 401–419.

Tversky, A., & Hutchinson, J. W. (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review*, *93*(11), 3–22.

van der Maaten, L. J. P. (2009). Learning a parametric embedding by preserving local structure. In *Proceedings of the twelfth international conference on artificial intelligence and statistics (AI-STATS), JMLR W&CP* (Vol. 5, pp. 384–391).

van der Maaten, L. J. P., & Hinton, G. E. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2431–2456.

van der Maaten, L. J. P., & Postma, E. O. (2010). Texton-based analysis of paintings. In *SPIE optical engineering and applications* (Vol. 7798-16).

Venna, J., Peltonen, J., Nybo, K., Aidos, H., & Kaski, S. (2010). Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, *11*, 451–490.

Villmann, T., & Haase, S. (2010). *Mathematical foundations of the generalization of t-SNE and SNE for arbitrary divergences* (Technical Report 02/2010). University of Applied Sciences Mittweida.

von Luxburg, U. (2010). Clustering stability: an overview. *Foundations and Trends in Machine Learning*, *2*(3), 235–274.

Weinberger, K. Q., Packer, B. D., & Saul, L. K. (2005). Nonlinear dimensionality reduction by semidefinite programming and kernel matrix factorization. In *Proceedings of the* 10*th international workshop on AI and statistics*. Barbados: Society for Artificial Intelligence and Statistics.

Yang, Z., King, I., Oja, E., & Xu, Z. (2010). Heavy-tailed symmetric stochastic neighbor embedding. In *Advances in neural information processing systems* (Vol. 22). Cambridge: MIT Press.