Robust Visualisation of Dynamic Text Collections: Measuring and Comparing Dimensionality Reduction Algorithms

Tim Repke
tim.repke@hpi.uni-potsdam.de
Hasso Plattner Institute
University of Potsdam, Germany

Ralf Krestel
ralf.krestel@hpi.uni-potsdam.de
Hasso Plattner Institute
University of Potsdam, Germany

ABSTRACT

Visualisations are supposed to provide intuitive ways to explore large document collections. State-of-the-art approaches usually transform high-dimensional representations of documents into 2-dimensional vectors using dimensionality reduction algorithms. These vectors are then placed into a landscape hopefully retaining semantic information regarding similarity from the high-dimensional representation. Traditionally, dimensionality reduction algorithms are developed with static collections in mind. However, many "real-world" document collections, such as news articles, scientific literature, patents, Wikipedia, or tweets, to name a few, grow and evolve over time. Visualising the temporal change of these collections poses various challenges for out-of-the-box dimensionality reduction algorithms.

In this paper, we propose strategies to adapt existing dimensionality reduction algorithms to incorporate change. These strategies ensure that landscapes at different intervals of the collection are robust with regard to spatio-temporal coherence. Furthermore, we propose metrics to measure the stability over time and compare several popular dimensionality reduction algorithms.

KEYWORDS

Dimensionality Reduction, Corpus Exploration, Visual Search

1 INTRODUCTION

Interaction with large document collections can take many forms. Keyword-based search interfaces are certainly the most popular among them. However, in more exploratory settings, users may find a 2-dimensional document landscape more compelling. They enable the user to visually explore a corpus in its entirety. There are several map-like visualisations for book corpora[1] [6] as well as philosophy [12] and physics literature [2]. These show the effectiveness of displaying individual documents in their global context, allowing the user to gain insights that would otherwise remain hidden. We focus on the aspect *where* a document should be placed on such a landscape, especially when the underlying document collection is growing. Most commonly, dimensionality reduction is applied to a high-dimensional representation of the original documents in the corpus. The most popular among them are tSNE and UMAP [10, 11]. However, these algorithms were developed with static data in mind. As the corpus grows over time, for example as new research is published or news events unfold, the landscape has to be updated. Updates to the landscape need to be coherent with their earlier versions so that users can reliably associate semantics to specific regions in their mental model of the data.

Dimensionality reduction algorithms typically start with an initial layout of the data in a two dimensional target space. This may either be an approximate layout, often using principle component analysis (PCA) or spectral embeddings, or sometimes a random distribution. The initialisation has a significant impact on the final layout after optimisation [15]. In this work, we compare several initialisation strategies in order to make landscape updates robust to undesired effects. Furthermore, we propose quantitative evaluation metrics to measure stability and layout quality.

Prior work has shown that consistent updates are possible under specific circumstances. Rauber et al. propose a dynamic tSNE approach by introducing a displacement penalty between layouts [17]. However, their focus was change in the representation of individual items in the dataset, not a growing corpus. LION-tSNE adds new data to the initial layout by calculating the position using k nearest neighbours of the original data [2, 4]. Poličar et al. extended tSNE by introducing batch processing to reduce the dimensionality of large datasets [15].

While batch processing assumes a relatively uniform distribution of all aspects of the data across batches, we assume the opposite. New topics may emerge over time and need to be fitted to earlier representations but may also require slight displacement of older data to make room. Similar issues are also considered in representation learning for dynamic word embeddings [1] or editable neural networks [22]. In both cases, unsupervised learning models are adapted with new objectives or to fit new data. In these scenarios, only pairwise relations need to be stable. For visualisation purposes however, the global rotation or distortion in landscape updates has to be limited.

There are several application that use dimensionality reduction to visualise document collections. Early work includes InVis [5, 13] and the work by Chen et al. who used probabilistic multidimensional projection models [3]. Repke et al. combined multiple objectives in their dimensionality reduction algorithm to jointly visualise

[1] http://galaxy.opensyllabus.org/
[2] https://github.com/ds3-nyu-archive/ds-dialect-map-ui

CHIIR 2021, March 15–19, 2021, online

ACM ISBN XXX-XX-XXXX-XXXX…$15.00
https://doi.org/XX.XXXX/XXXXXX.XXXXX

multi-faceted datasets, such as text with inherent network information [19] and demonstrated an exploration interface [18]. Schmidt et al. use stable random projections to visualise 15-million books from the Hathi Trust collection in a single explorable scatterplot [21]. Systems like Kyrix-S could augment this visualisation with the capability to aggregate information on lasso-selected data or query data [24]. For expert systems, certain characteristics of the layout of a landscape may be required. Pezzotti et al. proposed an algorithm to steer the dimensionality reduction algorithm during the iterative optimisation process [14].

In the following sections of this paper, we define strategies for initialising the layout process. We implemented them for several state-of-the-art dimensionality reduction algorithms. Finally, we evaluate the strategies using novel layout quality metrics and close with a qualitative discussion.

2 ROBUSTNESS THROUGH INITIALISATION

The initialisation has a significant impact on the final result of a dimensionality reduction algorithm. Thus, we hypothesise, that targeted initialisation strategies across visualisations for different intervals of a document collection leads to spatially coherent results. Obviously, some change to the position of earlier documents may be required to make room for emerging topics. This change however should be minimal. Furthermore, emerging topics should fit within the semantic structure of an existing layout. The development of new dimensionality reduction models is beyond the scope of this preliminary work. Thus, we only consider small adaptations of existing algorithms as baselines. For our experiments, we assume that the document collection is strictly growing and that the number of cumulated documents from earlier intervals to be larger than the number of documents in the next interval. To achieve coherence across intervals, we propose the following initialisation strategies.

The *naive approach* is to use the cumulated documents and the new documents as input to the dimensionality reduction algorithm and only fix parameters that could influence the layout, including the random state. Spectral embeddings or principle component analysis, commonly used for initialisation, both are deterministic when applied to the same data. In case the number of newly added documents is small enough compared to the number of already present documents, this may be a valid approach.

The *kNN approach* uses the layout of earlier documents to approximately place new documents within the existing layout. For each new document, we take k nearest neighbours in the high-dimensional space of earlier documents and calculate their average position in the two-dimensional space to place the new document. Alternatively the impact of outliers could be reduced by using the median or weighting the average by the inverse distances. We initialise the dimensionality reduction algorithms with these positions to optimise the two-dimensional layout.

Lastly, we also use *algorithm-specific approaches*. For example, UMAP [11] and Parametric-UMAP [20] provide a method to project previously unseen documents to the target space. Initially, this works similar to the kNN approach, but an internal encoder model refines the positions of new documents. OpenTSNE [16] implements advanced nearest neighbour methods to calculate affinities between data points in the high-dimensional space. These models can also be used as an initialisation in addition to the previously described kNN approach.

3 EXPERIMENTAL SETUP

In order to evaluate the different initialisation strategies described above, we implemented an abstract interface to implement all strategies for all algorithms. Prior work has shown that most dimensionality reduction algorithms are able to project previously unseen data into the two-dimensional space in one way or another. This only works when the new data is sampled from a similar distribution as the already seen data. This is obviously rarely the case for “real world” document collections. We therefore setup an experiment to isolate one particular scenario, namely the emergence of a new topic over time.

To this end, we simulate the dynamics of a growing document collection. We use the 20-newsgroup dataset [7] and completely hide documents from one category in the initial interval. Over time, only documents from the previously hidden category are added. This way, we simulate an emerging new aspect or topic in the document collection across intervals. The experimental results are averaged across multiple runs with different initially hidden categories.

The 20-newsgroup dataset contains around 18,000 texts assigned to one of 20 categories. We represent texts using 10,000-dimensional, tf.idf-weighted bag-of-words vectors. In interval one, we consider all documents from 19 categories but ignore all documents from the remaining category. For the second interval, we add 50 documents from the initially ignored category, for the third we add 200 more, and for the fourth we add the remaining documents from the class up to 1000 documents. This way, we simulate slow and rapid growth of the emerging topic over intervals and get an approximately uniform distribution over the categories.

We implemented interfaces for the most popular algorithms OpenTSNE [16], FItSNE [9], UMAP [11], Parametric-UMAP [20], and LargeVis [23]. The code is available for reproducibility on GitHub.[3]

4 EVALUATION METRICS

In this preliminary work, we propose strategies to adapt dimensionality reduction algorithms to produce coherent representations for document collections that grow and semantically evolve over time. An intuitively usable series of visualisations of the document collection should be relatively stable, such that an area on the landscape is always associated with a fixed semantic meaning. Furthermore, documents that existed in earlier versions of the landscape should remain near their original position for best usability. However, some displacement may be necessary to make space for emerging topics.

To evaluate the effectiveness of the proposed initialisation strategies with regard to these expectations, we measure 8 characteristics of the produced series of landscapes. Thereby we evaluate both, the landscapes themselves and the coherence within a series of landscapes. We base the individual evaluations on the work by Li et al. [8]. Among others, they used (1) the local Kullback-Leibler divergence (*L-KL*) to compare pairwise local distances in the original space and the landscape (lower is better), (2) a continuity score

[3] https://github.com/anon/anon (Repository URL redacted for double-blind review)

Table 1: Stability and Layout Quality Metrics, best score in bold, second best in *italics*

Algorithm	Strategy	L-KL	Acc	Cont	Trust	NMI	Spread	Overlap	Disp	Stab
OpenTSNE	naive	0.0529	0.6519	0.9048	*0.9150*	0.5234	0.2746	0.6136	0.3115	*0.7359*
OpenTSNE	specific	*0.0495*	**0.6560**	0.9093	**0.9192**	**0.5299**	0.2756	0.6414	*0.0738*	0.8559
FItSNE	naive	0.0539	0.6522	0.9058	0.9134	0.5239	0.2803	0.6192	0.2012	0.7937
FItSNE	kNN	0.0497	*0.6525*	0.9093	0.9106	*0.5247*	0.2700	0.5920	0.0880	**0.8632**
UMAP	naive	0.0517	0.5526	**0.9218**	0.8217	0.4499	*0.1876*	*0.1715*	0.1317	0.5848
UMAP	specific	0.0522	0.5576	*0.9214*	0.8221	0.4505	0.1882	0.1789	0.1229	0.4924
ParaUMAP	naive	0.0677	0.4206	0.8709	0.7193	0.3583	**0.1843**	*0.1715*	0.3254	0.2576
ParaUMAP	specific	0.0648	0.4457	0.8715	0.7386	0.3742	0.2010	**0.1625**	**0.0524**	0.8517
LargeVis	naive	**0.0379**	0.6428	0.9120	0.9055	0.5224	0.2278	0.5180	0.4351	0.4904

(*Cont*) which measures the overlap of point-wise k-neighbourhoods in both spaces, original and landscape (higher is better), (3) an accuracy score (*Acc*) that calculates how well the labels in the k-neighbourhood predict the label of each document on the landscape (higher is better), (4) and the trustworthiness (*Trust*) by calculating the number of points that appear in the k-neighbourhoods of the original space but not in the same neighbourhood on the landscape (higher is better).

However, these metrics, as well as others they used, only measure how well the landscape represents the high-dimensional space. Since our documents have category labels, we can also measure quality of their visual separation. To incorporate label information in the evaluation, we use (5) the average normalised mutual information (*NMI*) (higher is better), (6) the *Spread* of labels across the landscape (lower is better), (7) and the *Overlap* of Gaussian kernels that are fit to documents of each category (lower is better).

Furthermore, we measure the coherence in a series of landscapes by (8) the average normalised displacement (*Disp*) of data points from one landscape to the next (lower is better), (9) and the average overlap of Gaussian kernels (*Stab*) for each category between two landscapes to show how stable semantic areas are (higher is better).

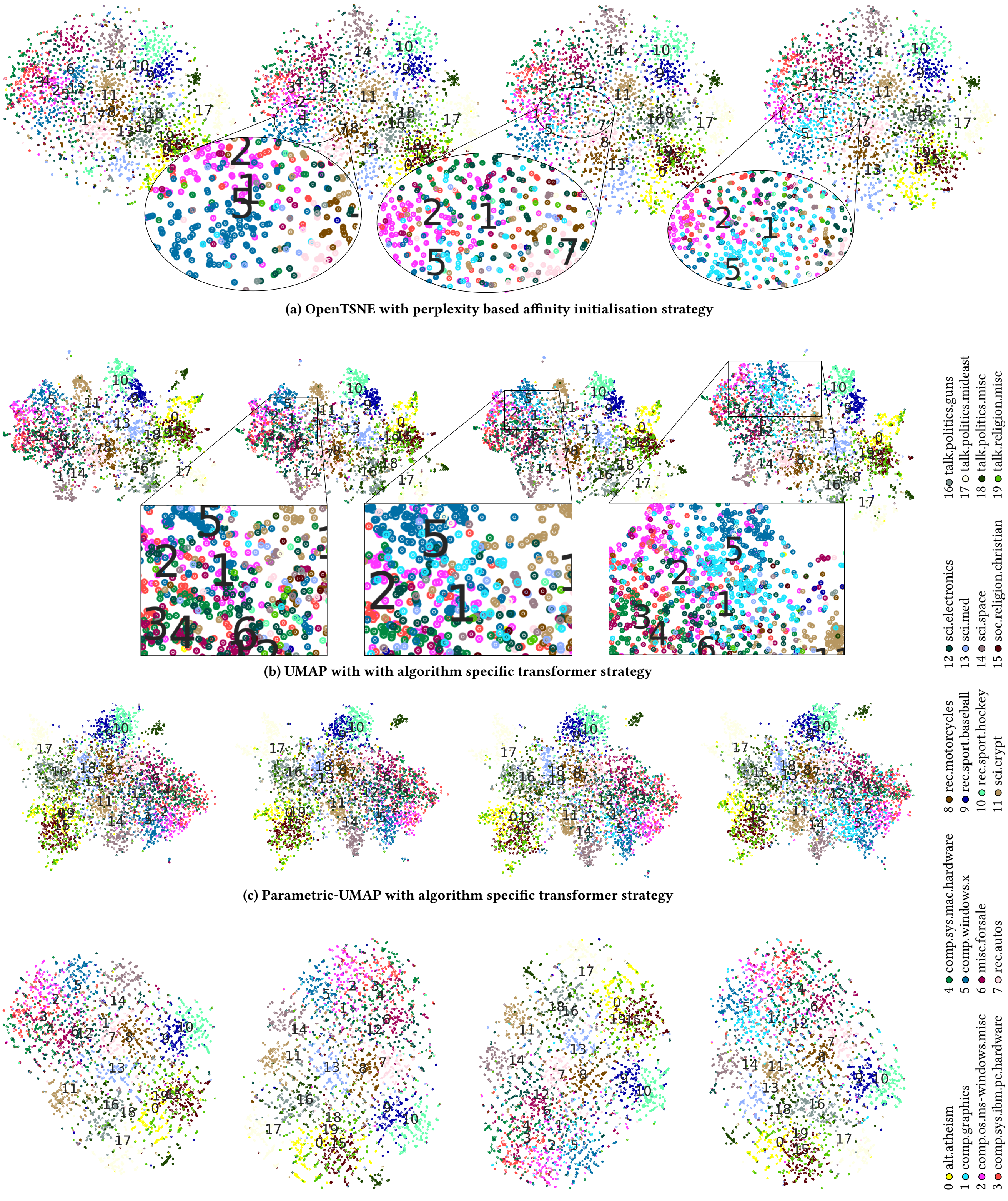
5 RESULTS

In this section we provide an overview of the results from our experiments. Table 1 contains the results of our quantitative evaluation. Due to space constraints, they are limited to the 20-newsgroup dataset and only provide a comparison of the naive approach for each dimensionality reduction algorithm and its best performing initialisation strategy. Based on these results and those not shown, none of the approaches and initialisation strategies appear to stand out. Comparing the tSNE-based and UMAP-based algorithms, we observe a trade-off in the results. UMAP-based approaches generally produce more stable layouts across intervals, as seen in the last four columns of the table. However, their ability to represent the original space decreases as a new category grows, which is not well separated from other documents in the existing landscapes. This becomes most apparent in the average NMI scores. Vice versa, tSNE-based approaches generally seem to incorporate documents from the emerging category better, while generally leading to less stable layouts, which is especially obvious when looking at all initialisation strategies. We also observed, that parameters for degrees of freedom or cluster separation appear to worsen all scores. We assume, that the separation amplifies the impact of documents, that appear at less ideal positions, given their category labels. Examples in Figure 1 were chosen based on most stable layouts, each corresponding to the non-naive strategy row in Table 1. These examples show, how documents from an emerging computer-related topic is embedded into the landscape around already existing similar topics. For the visualisation, we reduced the number of documents but still first add few documents from the initially hidden category and increase the growth rate. All algorithms lead to the least stability of the overall layout between the third and fourth interval. We can also see, that UMAP-based approaches tend to produce a more densely populated landscape than tSNE-based approaches.

6 CONCLUSIONS AND FUTURE WORK

In this paper we have shown the results of our comparison of different dimensionality reduction algoriithms to visualising document collections that grow and semantically evolve over time. To this end, we use the analogy of document landscapes for intuitive exploration of the global and local structure of the underlying document collection. These document landscapes need to be updated to incorporate new documents of the growing corpus. After each update, or even across several updates, users should still be able to recognise a familiar global placement of semantic regions. We compared several state-of-the-art dimensionality reduction algorithms and the influence of different initialisation strategies to achieve stable updates across intervals. We have shown that even simple strategies already improve these objectives over a naive approach. However, our preliminary findings also show that there is still room for improvement. For example, documents on emerging topics do not fit well in the overall structure and are hard to distinguish from older documents.

In future work, we will focus on that issue in particular and develop a cross-interval loss, so that the optimisation process does not solely rely on a good initialisation. In the context of an information retrieval and exploration system, the layout would also benefit from a better use of the space on a rectangular canvas. Furthermore, another interesting direction for future work is to update underlying document embedding models instead of only using a static vector space representation of the documents.



(a) OpenTSNE with perplexity based affinity initialisation strategy

(b) UMAP with with algorithm specific transformer strategy

(c) Parametric-UMAP with algorithm specific transformer strategy

(d) LargeVis with naive stragey (no cross-interval initialisation)

Figure 1: Scatterplots of selected results over four time intervals; 400 posts from each category; category "1 – comp.graphics" is hidden first and then growing over time by first adding 10, then 50, and finally 140 posts in later intervals

REFERENCES

[1] Robert Bamler and Stephan Mandt. 2017. Dynamic Word Embeddings. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 380–389. http://proceedings.mlr.press/v70/bamler17a.html

[2] Andrey Boytsov, François Fouquet, Thomas Hartmann, and Yves Le Traon. 2017. Visualizing and Exploring Dynamic High-Dimensional Datasets with LION-tSNE. *CoRR* abs/1708.04983 (2017), 44 pages. arXiv:1708.04983 http://arxiv.org/abs/1708.04983

[3] Yanhua Chen, Lijun Wang, Ming Dong, and Jing Hua. 2009. Exemplar-based visualization of large document corpus. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 15, 6 (2009), 1161–1168.

[4] Bheekya Dharamsotu, K. Swarupa Rani, Salman Abdul Moiz, and C. Raghavendra Rao. 2019. k-NN Sampling for Visualization of Dynamic Data Using LION-tSNE. In *26th IEEE International Conference on High Performance Computing, Data, and Analytics, HiPC 2019, Hyderabad, India, December 17-20, 2019*. IEEE, 63–72. https://doi.org/10.1109/HiPC.2019.00019

[5] Matthew W. Johnson, Michael Eagle, and Tiffany Barnes. 2013. InVis: An Interactive Visualization Tool for Exploring Interaction Networks. In *Proceedings of the 6th International Conference on Educational Data Mining, Memphis, Tennessee, USA, July 6-9, 2013*, Sidney K. D'Mello, Rafael A. Calvo, and Andrew Olney (Eds.). International Educational Data Mining Society, 82–89. http://www.educationaldatamining.org/EDM2013/papers/rn_paper_14.pdf

[6] Dmitry Kobak, George Linderman, Stefan Steinerberger, Yuval Kluger, and Philipp Berens. 2019. Heavy-tailed kernels reveal a finer cluster structure in t-SNE visualisations. In *Proceedings of the European Conference on Machine Learning (ECML)*. Springer-Verlag, Heidelberg, Germany, 11 pages.

[7] Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings*. Elsevier, 331–339.

[8] Stan Z. Li, Zelin Zhang, and Lirong Wu. 2020. Markov-Lipschitz Deep Learning. *CoRR* abs/2006.08256 (2020), 18 pages.

[9] George C Linderman, Manas Rachh, Jeremy G Hoskins, Stefan Steinerberger, and Yuval Kluger. 2019. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature methods* 16, 3 (2019), 243.

[10] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)* 9 (2008), 2579–2605.

[11] Leland McInnes and John Healy. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *CoRR* abs/1802.03426 (2018). arXiv:1802.03426 http://arxiv.org/abs/1802.03426

[12] Maximilian Noichl. 2019. Modeling the structure of recent philosophy. *Synthese* (2019), 1–12.

[13] Daniel Paurat and Thomas Gärtner. 2013. InVis: A Tool for Interactive Visual Data Analysis. In *Proceedings of the European Conference on Machine Learning (ECML) (Lecture Notes in Computer Science, Vol. 8190)*. Springer, 672–676.

[14] Nicola Pezzotti, Boudewijn PF Lelieveldt, Laurens van der Maaten, Thomas Höllt, Elmar Eisemann, and Anna Vilanova. 2017. Approximated and user steerable tSNE for progressive visual analytics. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 23, 7 (2017), 1739–1752.

[15] Pavlin G. Poličar, Martin Strazar, and Blaz Zupan. 2019. Embedding to Reference t-SNE Space Addresses Batch Effects in Single-Cell Classification. In *Discovery Science - International Conference DS (Lecture Notes in Computer Science, Vol. 11828)*. Springer, 246–260.

[16] Pavlin G. Poličar, Martin Stražar, and Blaž Zupan. 2019. openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. *bioRxiv* (2019), 2 pages.

[17] Paulo E. Rauber, Alexandre X. Falcão, and Alexandru C. Telea. 2016. Visualizing Time-Dependent Data Using Dynamic t-SNE. In *Eurographics Conference on Visualization, EuroVis 2016, Short Papers, Groningen, The Netherlands, 6-10 June 2016*, Enrico Bertini, Niklas Elmqvist, and Thomas Wischgoll (Eds.). Eurographics Association, 73–77. https://doi.org/10.2312/eurovisshort.20161164

[18] Tim Repke and Ralf Krestel. 2020. Exploration Interface for Jointly Visualised Text and Graph Data. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*. ACM Press, 73–74.

[19] Tim Repke and Ralf Krestel. 2020. Visualising Large Document Collections by Jointly Modeling Text and Network Structure. In *Proceedings of the Joint Conference on Digital Libraries (JCDL)*. ACM Press, 279–288.

[20] Tim Sainburg, Leland McInnes, and Timothy Q. Gentner. 2020. Parametric UMAP: learning embeddings with deep neural networks for representation and semi-supervised learning. *CoRR* abs/2009.12981 (2020), 23 pages.

[21] Benjamin Schmidt. 2017. Stable Random Projection: Minimal, universal dimensionality reduction for library-scale data. In *International Conference of the Alliance of Digital Humanities Organizations (DH)*. Alliance of Digital Humanities Organizations (ADHO), 3 pages.

[22] Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitriy Pyrkin, Sergei Popov, and Artem Babenko. 2020. Editable Neural Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net, 12 pages.

[23] Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. 2016. Visualizing large-scale and high-dimensional data. In *Proceedings of the International World Wide Web Conference (WWW)*. ACM Press, Geneva, Switzerland, 287–297.

[24] Wenbo Tao, Xiaoyu Liu, Yedi Wang, Leilani Battle, Çağatay Demiralp, Remco Chang, and Michael Stonebraker. 2019. Kyrix: Interactive pan/zoom visualizations at scale. *Computer Graphics Forum* 38, 3 (2019), 529–540.