Visualising Large Document Collections by Jointly Modeling Text and Network Structure

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

PRE-PRINT VERSION

Many large text collections exhibit graph structures, either inherent to the content itself or encoded in the metadata of the individual documents. Example graphs extracted from document collections are co-author networks, citation networks, or named-entity-cooccurrence networks. Furthermore, social networks can be extracted from email corpora, tweets, or social media. When it comes to visualising these large corpora, either the textual content or the network graph are used.

In this paper, we propose to incorporate both, text and graph, to not only visualise the semantic information encoded in the documents' content but also the relationships expressed by the inherent network structure. To this end, we introduce a novel algorithm based on multi-objective optimisation to jointly position embedded documents and graph nodes in a two-dimensional landscape. We illustrate the effectiveness of our approach with real-world datasets and show that we can capture the semantics of large document collections better than other visualisations based on either the content or the network information.

CCS CONCEPTS

• **Information systems** → Digital libraries and archives; *Document collection models*; • **Computing methodologies** → **Semantic networks**; • **Human-centered computing** → *Visualization theory, concepts and paradigms.*

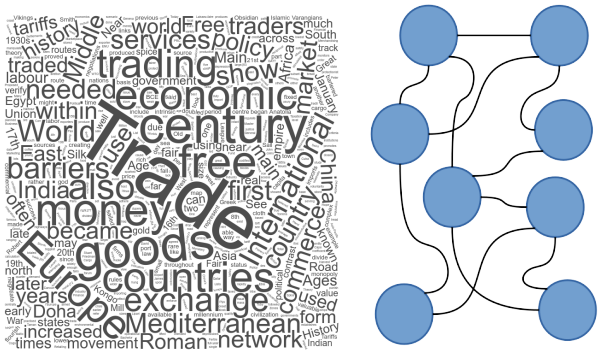
KEYWORDS

Corpus Exploration, Dimensionality Reduction, Corpus Visualisation

ACM Reference Format:
Tim Repke and Ralf Krestel. 2020. Visualising Large Document Collections by Jointly Modeling Text and Network Structure. In *Proceedings of the 20th Joint Conference on Digital Libraries (JCDL). June 19–23, 2020. Wuhan, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/00.0000/0000000.0000000

JCDL '20, June 19–23, 2020, Wuhan, China

ACM ISBN 000-0-0000-0000-0/00/00…$0.00
https://doi.org/00.0000/0000000.0000000



(a) Content-based word cloud **(b) Metadata-based network**

Figure 1: Duality of text collections.

1 INTRODUCTION

Exploring large document collections is a cumbersome, but necessary task to gain an overview or to find interesting, serendipitous information. Depending on the collection, the exploration either focuses on the content, for example by using topic modeling methods to get an overview, or on the network formed by the connections of documents among each other. Most digital library collections exhibit this duality; they can be represented as *text* or *network*. Figure 1 exemplifies how the two representations can be visualised, e.g. using word clouds and graphs. The duality is most apparent in collections of web pages, where links connect the pages with each other to form the Web graph. But is is also prevalent in email collections or corpora originating from communication in social networks, such as chats, blogs, or tweets. Often, analysing the communication network is more revealing than focusing on the content. While these are some examples of document collections that exhibit explicit network structure, most document collections can be enriched with network structure by extracting information from the content or by analysing the documents' metadata. For example, bibliometrics makes heavy use of both types of information: content of documents (research publications, patents, etc.) and co-author and (co-)citation networks. Visualising corpora is inevitable to analyse or explore the collections. But usually either the content or the network structure is neglected, missing out on important relations and insights about the document collection at hand.

In more heterogeneous data collections, exploration or getting an overview of datasets is insurmountable with current tools. The

sheer amount of documents prohibits simple visualisations of networks or meaningful keyword-driven summaries of the textual content. Examples of these extremly difficult cases are in the context of data-driven journalistm, computational forensics, or auditing. Data-driven journalism [8] often has to deal with leaked, unstructured, very heterogeneous data, e.g. in the context of the Panama Papers, where journalists needed to untangle and order huge amounts of information, search entities, and visualise found patterns [5]. Similar datasets are of interest in the context of computational forensics [14]. Auditing firms and law enforcement need to sift through huge amounts of data to gather evidence of criminal activity, often involving communication networks and documents [22]. Users investigating such data want to be able to quickly gain an overview of its entirety, since the large amount of heterogeneous data renders experts' investigations by hand infeasible. Computer-aided exploration tools can support their work to identify irregularities, inappropriate content, or suspicious patterns. Current tools[1] lack sufficient semantic support, for example by incorporating document embeddings [30] and the ability to combine text and network information intuitively.

We propose *MODiR*, a scalable **M**ulti-**O**bjective **Di**mensionality **R**eduction algorithm, and show how it can be used to generate an overview of entire document collections with inherent network information in a single interactive visualisation. Special graph databases enable the efficient storage of large relationship networks and provide interfaces to query or analyse the data. However, without prior knowledge, it is practically impossible to gain an overview or quick insights into global network structures. Although traditional node-link visualisations of a network can provide this overview, all semantic information from associated textual content is lost completely.

Technically, our goal is to combine network layouts with dimensionality reduction of high-dimensional semantic embedding spaces. Giving an overview over latent structures and topics in one visualisation may significantly improve the exploration of a corpus by users unfamiliar with the domain and terminology. This means, we have to integrate multiple aspects of the documents, namely the semantics of the textual content and the relations and connections inherent to the collection, into a single visualisation. The challenge is to provide an intuitive, two-dimensional representation of both the network and the text, while balancing potentially contradicting objectives of these representations.

In contrast to existing dimensionality reduction methods, such as tSNE [28], we propose a novel approach to transform high-dimensional data into two dimensions while *optimising multiple constraints* simultaneously to ensure an optimal layout of semantic information extracted from text and the associated network. To minimise the computational complexity that would come from a naive combination of network drawing and dimensionality reduction algorithms, we formally use the notion of a hypergraph. In this way, we are able to move repeated expensive computations from the iterative document-centred optimisation to a preprocessing step that constructs the hypergraph. We use real-world document collections from different domains to demonstrate the effectiveness and flexibility of our approach. *MODiR*-generated representations are compared to a series of baselines and state-of-the-art visualisation and dimensionality reduction methods. We further show that our integrated view of these document collections is superior to approaches focusing on text-only or network-only information when computing their visualisations.

[1] e.g. https://www.nuix.com/ or https://linkurio.us/

2 RELATED WORK

With *MODiR* we bridge the gap between text and network visualisation by jointly reducing the dimensionality of the input data. Therefore we subdivided this part into three sections to highlight related work in the areas of text visualisation, representation learning, as well as dimensionality reduction. Other work that tries to jointly model text and networks but without dimensionality reduction and without a focus on visualisation is *LINE* [44]. They generate information networks consisting of different types of nodes, e.g. words from document content and authors from document metadata. Another tool that investigates combining graph structure with textual elements is *VOSviewer* [46]. They construct and visualise bibliographic networks that provide a multi-view interface to explore and filter keywords and network aspects of such datasets. In our work we go beyond building a network from textual data but instead project the textual data into a latent space.

Document visualisation aims to visualise the textual content, such that users gain quick insights into topics, latent phrases, or trends. Tiara [47] extracts topics and derives time-sensitive keywords to depict evolving subjects over time as stacked plots. Another line of work projects documents into a latent space, for example by using topic models or embeddings: Creating scatter-plots of embedded documents of a large corpus may result in a very dense and unclear layout, so Chen et al. [7] developed an algorithm to reduce over-full visualisations by picking representative documents. A different approach is taken by Fortuna et al. [13], who do not show documents directly, but generate a heatmap of the populated canvas and overlay it with salient phrases at more densely populated areas from the underlying documents in that region. Friedl et al. [15] extend that concept by drawing clear lines between regions and colouring them. They also add edges between salient phrases based on co-occurrences in the texts. A map analogy can be used to visualise the contents of documents by embedding them into a high dimensional semantic space [25] and projecting it on a two-dimensional canvas as a *document landscape*. Most recently *Cartograph* [41] was proposed, which is visually very similar to previous approaches, but pre-renders information at different resolution and uses a tiling server with (geographic) map technology to deliver responsive interactions with the document landscape. Regions are coloured based on underlying ontologies from a knowledge-base.

Networks are traditionally visualised using so-called node-link graphs. This way, any additional information related to nodes and edges are lost. The layout of nodes usually follows a force-based analogy first proposed by Fruchterman and Reingold [16]. Newer approaches optimise the computational complexity and include local metrics to better represent inherent structures as for example *ForceAtlas2* [21], which is the default network algorithm for the network visualisation tool *Gephi*. Besides these traditional systems, more exotic approaches use the metaphor of geographical maps [33]

to visualise networks, for example using topology to reflect connectivity of densely connected social communities. In order to highlight how relationships form and change based on the interactions, the metaphor of a growing tree can be used (ContactTrees [38]). Although this reflects temporal aspects of dynamic networks well, it focuses on one person as the root, thus an overview of the entire network is not possible. CactusTrees [9], on the other hand, represent hierarchical structures with the goal of untangling overlaid bundles of intersecting edges, making distant connections more apparent. Usually, a communication network has many nodes and overlapping connections already, so Yang et al. [48] rather focus on discovering overlapping cores to improve the identification of community boundaries to highlight global latent structures. Similarly, Gronemann et al. [17] use the metaphor of islands and hills to visualise clustered graphs, making densely connected communities clearly noticeable. But, the edges are bundled and follow valleys of the resulting topology, thus making relationships between other communities hard to follow. MapSets [11] assume a graph that was laid out using embeddings reflecting communities. An algorithm then draws regions around clusters of nodes, such that the bounding shapes are contiguous and non-overlapping, but yet abstract. Another approach to visualise networks at full scale is to aggregate nodes based on their spatial distribution and thereby allowing for a simple exploration with contour lines and heatmap overlays to emphasise latent structures as proposed by Hildenbrand et al. [19].

The text and network visualisation methods discussed above primarily use structural properties of the data to generate their layout. Although we focus on the visualisation of text data with inherent graph information, *MODiR* can work with arbitrary kinds of data. Our model only requires a way to project the data into a high-dimensional Euclidean vector space so that the distance between two points can be interpreted as their (semantic) similarity. Traditionally, text can be represented as bag-of-words vector that optionally is weighted by respective tf-idf scores. In recent years, embeddings became more popular as they conserve semantic meaning in their vector representation. Mikolov et al. [30] introduced neural architectures to learn high-dimensional vector representations for words and paragraphs [25]. Similar methods are used to learn representations for nodes in a network based on either the structural neighbourhood [12] or additional heterogeneous information [6, 27]. Schlötter et al. [40] attempted to learn joint representations of network structure and document contents but saw no improvement over conventional models in a series of classification tasks. We only use the structural information of the network for better control over fine-grained adjustments in our layout algorithm. Literature on graph embeddings is sometimes qualitatively evaluated by visualising the dimensionality reduced embedding space [49]. More specifically, Hamilton et al. [18] have shown that simple document and word embeddings can be enriched by using graph convolutions over a network of co-occurrence statistics. In this work however, we refrain from using network embeddings, as it allows us to better utilise the network characteristics.

The goal of dimensionality reduction is to represent high-dimensional data in a low-dimensional space while preserving the characteristics of the original data as sound as possible. A very common application of dimensionality reduction is to project high-dimensional data into two dimensions for the purpose of visual interpretation. Generally, these methods follow one of three mathematical models. *Linear* models, such as Principal Component Analysis (PCA) [34] can be calculated very efficiently and have proven to reduce input spaces to improve the performance of downstream tasks. Thus, they are often indirectly used for feature extraction. Although reductions to two dimensions for visualisations are appropriate for quick initial data exploration, other approaches are able to better preserve data characteristics in two dimensions. For example, the *non-linear* Sammon mapping [39] tries to preserve the structure of inter-point distances in high-dimensional space in low-dimensional space. The resulting visualisations are generally better then PCA to show relatedness of individual data points. Lastly, there are *probabilistic* models like Stochastic Neighbour Embeddings (SNE) [20]. They are similar to a Sammon mapping in that they use inter-point distances but model these distances as probability distributions. The t-distributed SNE has proven to produce competitive results for visualising datasets while preserving characteristics [28], however its nondeterministic nature may produce greatly varying results. Recently, *FltSNE* was proposed, an optimisation of tSNE that significantly reduces the computational complexity [26]. Other newer dimensionality reduction algorithms like *LargeVis* [43] and *UMAP* [29] scale almost linearly by using efficient nearest neighbourhood approximations in the high-dimensional space and spectral embeddings to initialise positions of points in the low-dimensional space to reduce the number of fine-tuning iterations.

3 MULTI-OBJECTIVE DIMENSIONALITY REDUCTION

Visualisations of complex datasets are restricted to two or three dimensions for users to grasp the structure and patterns of the data. We integrate multiple kinds of information (i.e., documents and persons) into a joint visualisation as depicted on the far right in Figure 2, which we call *landscape*. This landscape consists of a base-layer containing all documents depicted as dots forming the *document landscape*; nodes and their connections are placed on top of this base-layer as circles connected by lines forming the *graph layer*. In this section, we propose the *MODiR* algorithm which integrates multiple objectives during the layout process to find an overall good fit of the data within the different layers. Our approach is derived from state-of-the-art methods for drawing either the network layer or the document landscape. We formally model the data as part of a hypergraph, which we abstractly depict on the left in Figure 2. This allows for a more simple implementation of the algorithm and easier data structures that operate on (cached) sets as opposed to traversing a "normal" graph structure.

We assume that documents are given as high-dimensional vectors and entities are linked among one another and to the documents. These links are used as restrictions during the multi-objective dimensionality reduction of document vectors. Let $\boldsymbol{x}^{(i)} \in \mathbb{X} \subset \mathbb{R}^d$ be the set of n documents in their d-dimensional representation and $\boldsymbol{y}^{(i)} \in \mathbb{Y} \subset \mathbb{R}^2$ the respective positions on the document landscape. Let $\mathcal{H}(\mathcal{V}, \mathcal{E})$ be a hypergraph based on the network information inferred from the document corpus, with vertices $\mathcal{V} = \mathbb{X} \bigcup \mathbb{P}$, where $\mathbb{X}$ are the documents and $p_i \in \mathbb{P}$ are the entities in the network and hyperedges $e_k \in \mathcal{E}$ describing the relation between documents

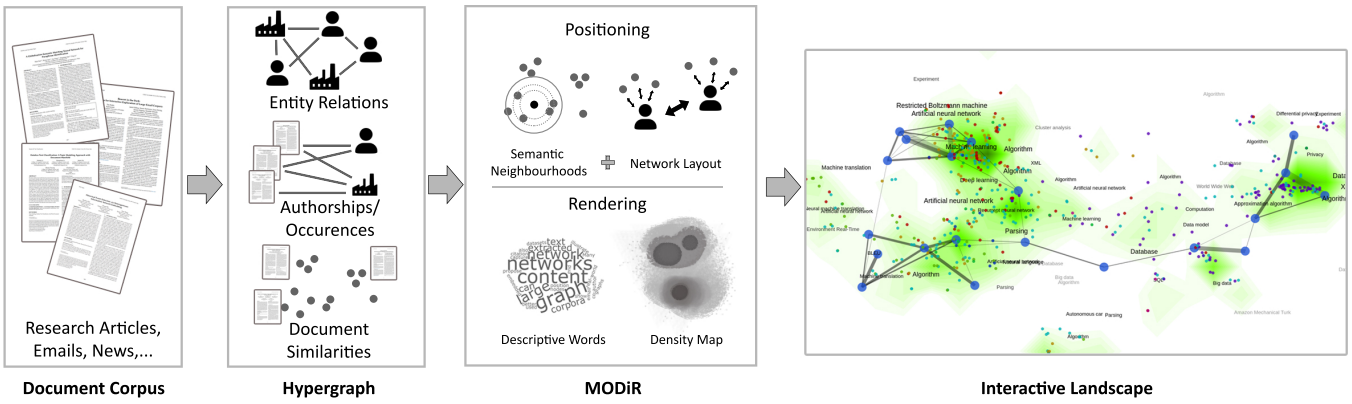


Figure 2: Overview of *MODiR* for joint visualisation of research articles with co-authorship networks, email corpora, and more

and entities. For each pair of entities $p_m, p_n \in \mathbb{P}$ that are connected in the context of documents $\boldsymbol{x}^{(i)}, \ldots \in \mathbb{X}$, there is a hyperedge $e_k = \{p_m, p_n, \boldsymbol{x}^{(i)}, \ldots\}$. Analogously, the same definition applies to $\mathbb{Y}$. Further, $\mathcal{H}^{\mathbb{Y}}$ or $\mathcal{H}^{\mathbb{X}}$ is used to explicitly state the respective document representation used. The position in the graph layer $\pi : \mathbb{P} \rightarrow \mathbb{R}^2$ of an entity p_m is defined as

$$\pi(p_m; \mathcal{H}^{\mathbb{Y}}) = \frac{1}{N_{p_m}} \sum_{e_k \in \mathcal{E}_{p_m}} \sum_{\boldsymbol{y}^{(i)} \in e_k \setminus \mathbb{P}} \boldsymbol{y}^{(i)}, \tag{1}$$

where $\mathcal{E}_{p_m} \subset \mathcal{H}^{\mathbb{Y}}$ is the set of hyperedges containing p_m and N_{p_m} is the number of documents p_m is associated with.[2] This effectively places an entity at the centre of its respective documents. More elaborate methods like a density-based weighted average are also applicable to mitigate the influence of outliers. For simplicity we will abbreviate $\pi(p_m; \mathcal{H}^{\mathbb{Y}})$ as π_m.

Let $\psi : \mathbb{X} \rightarrow \mathbb{Y}$ be the projection $\psi(\boldsymbol{x}^{(i)}; \boldsymbol{W}) = \boldsymbol{W}_{i,:} = \boldsymbol{y}^{(i)}$, where $\boldsymbol{W} \in \mathbb{R}^{2 \times n}$ is the projection matrix leant by *MODiR* based on multiple objectives $\varphi_{\{1,2,3\}}$ using gradient descend, as defined later in this section. The objectives are weighted by manually set parameters $\theta_{\{1,2,3\}}$ to balance the effects that favour principles focused on either the graph layer or the document landscape, as they may contradict one another. Given a high-dimensional hypergraph $\mathcal{H}^{\mathbb{X}}$, the matrix $\boldsymbol{W}$, and a entity projection π, we define the resulting multi-objective dimensionality reduction function as

$$\Psi(\mathcal{H}^{\mathbb{X}}, \boldsymbol{W}, \pi) = \mathcal{H}^{\mathbb{Y}}.$$

We summarise the most important definitions in Table 1.

In the following paragraphs, we will formally introduce *MODiR*'s objectives. *Objectives (1) and (2)* are inspired by tSNE and use the neighbourhood context of documents in $\mathbb{X}$ to position similar documents near one another and unrelated ones further apart in $\mathbb{Y}$. *Objective (3)* attracts documents based on co-occurrence in hyperedges so that the resulting π_m will be closer if they are well connected in the graph. This third objective also implicitly brings documents closer to their respective entities.

[2] $N_{p_m} := \left|\{\boldsymbol{x}^{(i)} \in \mathbb{X} | \exists e_k \in \mathcal{E} : \boldsymbol{x}^{(i)} \in e_k \wedge p_m \in e_k\}\right|$

Table 1: Overview of Symbols

Symbol	Description
$\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}$	Document vector and its position on the landscape
p_i, π_i	Entity in the graph and its position on the landscape
φ_1	Objective to pairwise attract similar documents
φ_2	Objective to pairwise repel dissimilar documents
φ_3	Objective to attract pairs of documents and entities
$\theta_{\{1,2,3\}}$	Weights for influence of objectives on Φ
$\mathbb{X}^{k,\boldsymbol{x}^{(i)}}$	Semantic neighbourhood of $\boldsymbol{x}^{(i)}$ with size k
$\bar{\mathbb{X}}^{l,\boldsymbol{x}^{(i)}}$	Non-similar neighbourhood of $\boldsymbol{x}^{(i)}$ with size l
$\mathcal{E}^{\mathbb{X}}_{\boldsymbol{x}^{(i)}}$	Set of documents connected to $\boldsymbol{x}^{(i)}$ via any entity; sampled down to size s

Objective (1): Similar documents are near one another. Semantically similar documents should be closer on the document landscape and dissimilar ones further apart. To measure the semantic similarity of documents, Maaten et al. [28] used a naïve bag-of-words representation. Although tSNE preserves the inherent semantic structure in two-dimensional representations from these sparse vectors [35], we opted to use document embeddings. This has the advantage that, when only part of the data is visualised, the embedding model can still be trained on a larger set of documents and thus retain the additional information. Objective (1) is inspired by the efficient usage of context words in word2vec [30]. Corresponding to the skip-gram model, we define the context $\mathbb{X}^{k,\boldsymbol{x}^{(i)}} \subset \mathbb{X}$ of a document $\boldsymbol{x}^{(i)}$ by its k nearest neighbours in the embedding space. The first objective is defined as

$$\varphi_1(x^{(i)}) = \sigma\Big(\sum_{\boldsymbol{x}^{(j)} \in \mathbb{X}^{k,\boldsymbol{x}^{(i)}}} \|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\| - \|\boldsymbol{y}^{(i)} - \boldsymbol{y}^{(j)}\| \Big) \tag{2}$$

with σ being the sigmoid function and $\|\cdot\|$ the Euclidean norm. Distances are normalised based on the context to make them comparable between the high-dimensional and two-dimensional space and rescaled by the sigmoid.

Objective (2): Dissimilar documents are apart from one another. The optimal solution to the previously defined objective would be to

project all documents onto the same point on the two-dimensional canvas. In order to counteract that, we introduce negative examples for each pair of context documents. We do so by sampling a set of l documents that are not in the k neighbourhood of $\boldsymbol{x}^{(i)}$. Let $\bar{\mathbb{X}}^{l,\boldsymbol{x}^{(i)}} \subset \mathbb{X} \setminus \mathbb{X}^{k,\boldsymbol{x}^{(i)}}$ be the set of negative samples for $\boldsymbol{x}^{(i)}$, then the second objective is defined as

$$\varphi_2(\boldsymbol{x}^{(i)}) = -\sigma\Big(\sum_{\boldsymbol{x}^{(j)} \in \bar{\mathbb{X}}^{l,\boldsymbol{x}^{(i)}}} \|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\| - \|\boldsymbol{y}^{(i)} - \boldsymbol{y}^{(j)}\| \Big). \quad (3)$$

This objective prevents crowding on the centre of the landscape and helps to better preserve the global structure.

Objective (3): Connected entities are near one another and their documents. This object serves two purposes: All documents $\boldsymbol{y}^{(i)}$ associated with an entity p_m are placed near its π_m position in the graph layer and two entities π_m and π_n are forced near one another if they are connected.

Let $\mathcal{E}_{\boldsymbol{y}^{(i)}} \subset \mathcal{E}$ be the set of hyperedges in the hypergraph $\mathcal{H}$ containing the document $\boldsymbol{y}^{(i)}$ and $\mathcal{E}^{\mathbb{Y}}_{\boldsymbol{y}^{(i)}} = \bigcup_{e_k \in \mathcal{E}_{\boldsymbol{y}^{(i)}}} e_k \setminus \mathbb{P}$ all documents that are linked to $\boldsymbol{y}^{(i)}$ through an entity, then the third objective is defined as

$$\varphi_3(\boldsymbol{y}^{(i)}) = \sigma\Big(\sum_{\boldsymbol{y}^{(j)} \in \mathcal{E}^{\mathbb{Y}}_{\boldsymbol{y}^{(i)}}} \|\boldsymbol{y}^{(i)} - \boldsymbol{y}^{(j)}\| \Big), \quad (4)$$

which, when minimised, attracts documents that are related through entities. This has two implicit effects: An entity p_m gets closer to its documents as they are attracted to π_m without having to explicitly compute this position using Equation (1). Also, related entities p_m, p_n are attracted to one another since they appear in the same hyperedges. The computational complexity of this objective is strongly related to the connectedness of entities in the graph. For dense graphs, we propose a heuristic by only using a subset of s documents from the context $\mathcal{E}^{\mathbb{Y}}_{\boldsymbol{y}^{(i)}}$ of $\boldsymbol{y}^{(i)}$. An objective modelling a repulsive force as in force-directed graph layouts is not needed as the first two objectives $\varphi_{\{1,2\}}$ provide enough counteracting force.

Algorithm. The positions of entities and documents on the landscape are calculated using the previously defined objectives as follows. First, we construct the hypergraph $\mathcal{H}^{\mathbb{X}}$ with document contexts including the set of k-neighbourhoods $\mathbb{X}^{k,\boldsymbol{x}^{(i)}}$. Relevant pairwise distances can be stored in an adjacency matrix so reduce computational overhead in Equations 2 and 3. For more efficient training, the randomly sampled l negative neighbourhoods $\bar{\mathbb{X}}^{l,\boldsymbol{x}^{(i)}}$ can be prepared ahead of time and then only masked during later. The s-neighbourhoods for entities in Equation (4) $\mathcal{E}^{\mathbb{Y}}_{\boldsymbol{y}^{(i)}}$ can only be prepared with references, as $\mathbb{Y}_{\boldsymbol{y}^{(i)}}$ updates with each iteration. We designed the algorithm to move as much repetitive computations to pre-processing ahead of time or each epoch. Creating these sets is very efficient using Hierarchical Navigable Small World graphs (HNSW) for approximate nearest neighbour search [3]. Overall we are able to reduce the pre-processing complexity to $O(n \log n)$ and for each iteration $O(kln)$, with $k, l \ll n$ near linear. After generating the context sets, we use gradient descend to update the projection matrix $\boldsymbol{W}$ (rows are $\boldsymbol{y}^{(i)}$) with learning rate η reducing the overall error Φ as defined by

$$\Phi(x_i) = \theta_1 \varphi_1(\boldsymbol{x}^{(i)}) + \theta_2 \varphi_2(\boldsymbol{x}^{(i)}) + \theta_3 \varphi_3(\boldsymbol{x}^{(i)}). \quad (5)$$

Selecting appropriate values for the hyperparameters k, l, s, and $\theta_{\{1,2,3\}}$ is critical to produce meaningful results. We found $l = k$ in all experiments to produce the best results as this way for every similar document the model has one dissimilar document to compare. Inspired by tSNE [28], we limit hyperparameters by setting k and s dynamically for each document based on a user-defined perplexity. With these adaptations, the only parameters to be set are the perplexity β that roughly determines the context size, the learning rate η, and the objective weights, which can often stay at a default setting. A reference implementation including a modular processing pipeline for different datasets, approaches, and experiments is available on GitHub[3].

Table 2: Number of documents, entities, and their connections in filtered datasets used in this paper

Dataset	# Documents	# Nodes	# Edges
AMiner (AM)	49,670	56,449	110,146
SemanticScholar (S2)	170,098	183,198	701,442
SmallScholar (S2b)	489	24	39
Enron (ENR)	189,437	32,353	950,100
News (NEW)	3,734	2,944	5,240

4 EXPERIMENTS

Our approach can be used in a variety of different scenarios. Communication datasets, such as emails, are particularly interesting, since understanding this data or getting an overview of it necessitates the analysis and visualisation of both, content and meta-data. While there exists these kinds of document collection, e.g. the Enron corpus [23], they typically lack ground truth for evaluation purposes. Another type of document collections is more accessible regarding evaluation: research publications and their co-authorship network. Therefore we focus our experiments on collections of scientific articles and how they can be visualised using their content and information about co-authorship. Results of dimensionality reduction can be subjective, so as in prior work on dimensionality reduction [28, 29, 41], we will qualitatively compare our approach to a variety of baselines but in addition we will provide a few quantitative experiments as well. To the best of our knowledge, there are no algorithms that use multiple objectives for dimensionality reduction of high-dimensional data. Popular approaches for traditional dimensionality reduction are tSNE and PCA. As baselines, we use the original optimised implementation of tSNE[4] written in C as provided by the authors.

4.1 Datasets

The motivation for this paper is to visualise inherent network structure along with their respective text documents for exploring and

[3] https://github.com/redacted/redacted (link will be part of camera ready version)
[4] https://lvdmaaten.github.io/tsne/

understanding large document collections. We argue, that our approach is applicable to any given document collection with inherent graph structures, so we include a variety of examples for evaluation. We apply *MODiR* to the Enron corpus [23] which originally consists of around 600,000 messages belonging to 158 users and Quagga [37] to extract individual emails from quoted conversations, remove duplicates, extract additional correspondents from inline metadata, and try to combine the aliases of people. Assessing the quality of a given layout requires very specific domain knowledge including deep understanding of semantic structure across all documents and a close familiarity with entity relations. Due to the lack of a gold standard or domain knowledge on our side, so we consider additional sources. Thus we use named entities extracted from business news articles. From the corpus of 448,395 Bloomberg- and 106,519 Reuters news articles (NEW) published by Ding et al [10], we select those that contain the search term "commerzbank" as a central entity and consider co-occurrences of organisation entities extracted with AmbiverseNLU [42]. This results in a graph where almost all entities are connected to a single central entity that appears in all articles.

Academic co-authorship networks and their respective publications have well defined labels provided by venues or communities, so there are no ambiguities or additional annotations needed. We make use of two processed and publicly available corpora of research articles, the AMiner[5] network (AM) [45] published in 2008 with over two million papers by 1.7 million authors and the recently published Semantic Scholar[6] Open Corpus (S2) [1] with over 45 million articles. Both corpora cover a range of different scientific fields. Semantic Scholar for example integrates multiple data sources like DBLP and PubMed and mostly covers computer science, neuroscience, and biomedical research. Unlike DBLP however, S2 and AM not only contain bibliographic metadata, such as authors, date, venue, citations, but also abstracts to most articles, that we use to train document embeddings using the Doc2Vec model in Gensim [7]. Similar to Carvallari et al. [4] remove articles with missing information and limit to six communities that are aggregated by venues as listed in Table 3. This way we reduce the size and also remove clearly unrelated computer science articles and biomedical studies. For in depth comparisons we reduce the S2 dataset to 24 hand-picked authors, their co-authors, and their papers (S2b).

Note, that the characteristics of the networks differ greatly as the ratio between documents, nodes, and edges in Table 2 shows. In an email corpus, a larger number of documents is attributed to fewer nodes and the distribution has a high variance (some people write few emails, some a lot). In the academic corpora on the other hand, the number of documents per author is relatively low. Especially different is the news corpus, that contains one entity that is linked to all other entities and to all documents.

4.2 Hyperparameter Settings

For *MODiR*, the context sizes are the most important parameters. Generally, small numbers for k, l, s perform better. This is in line with our expectations, as each item $\boldsymbol{x}^{(i)}$ will also be in the context of its respective neighbours and will therefore amplify its attractive force. A large number for k for example will force all points towards the centre of the canvas or if even larger, produce random scatter as the gradients amplify. In our experiments we use $k = 10$, for datasets with a few thousand samples, k should usually be below l. We also found, that the negative context is best with $l = 20$ for all sizes.

Furthermore, we set both $\theta_1 = \theta_2 = 1.0$ for all experiments because the influence on selecting k, l is much larger. The graph context is also set to $s = 10$ (in our dataset the number of entities is close to the number of documents), the objective weight can be freely adjusted between around $0.8 \leq \theta_3 \leq 1.2$ to set the influence of the entity network. Similar to the semantic neighbourhoods in the first and second objective, the choice of s is significantly more influential than θ_3. Setting $\theta_1 = \theta_2 = 0$ to get a network-only layout would not work as the optimum would be placing all points in the middle as discussed earlier. However, it is possible to "turn off" the influence of the network information on the layout by setting $\theta_3 = 0$.

The speed of convergence depends on the learning rate η and thus dictates the number of maximum iterations. Early stopping with a threshold on the update rate could be implemented. Depending on the size of the dataset and a fixed learning rate of $\eta = 0.01$, *MODiR* generally converges after 10 to 200 iterations, for larger and more connected data it is advisable to use a higher learning rate in the first epoch for initialisation and then reducing it to very small updates. For better comparability, we use a constant number of iterations of $T = 100$. In our experiments using tSNE, we set the perplexity to $Perp(P_i) = 5$, $\theta = 0.5$ and run it for 1,000 iterations.

4.3 Quantitative Evaluation

As Maaten et al. [28] state, it is by definition impossible to fully represent the structure of intrinsically high-dimensional data, such as a set of document embeddings, in two dimensions. However, stochastic neighbour embeddings are able to capture intrinsic structures well in two dimensional representations [24]. To measure this capability, we compare the ability of k-means++ [2] to cluster the high- and two-dimensional space. We set the number of clusters to the number or research communities ($k = 6$) and calculate the percentage of of papers for each community per cluster. Therefore we assign each community to the cluster with most respective papers and make sure to use a clustering with an even distribution. Results are listed in Table 3 for tSNE, PCA, *MODiR*, and the original high dimensional embedding averaged over five runs. We see, that as expected due to topical overlap of communities, even original embeddings cannot be accurately clustered. Interestingly though, there seems to be a significant difference between AM and S2 although the sets of papers intersect, which we assume is due to the fact, that S2 is larger and additionally contains more recent papers. Although PCA often does not generate visualisations in which classes can be clearly distinguished, the clustering algorithm is still able to separate them with competitive results compared to tSNE and *MODiR*.

MODiR not only aims to produce a good document landscape, but also a good layout of the network layer. Graph layouts are well studied, thus we refer to related work on aesthetics [36] and

[5] https://aminer.org/billboard/aminernetwork
[6] https://api.semanticscholar.org/corpus/
[7] https://radimrehurek.com/gensim/; embedding size: 64 dimensions, vocabulary size: 20k tokens, trained for 500 epochs

Table 3: Selected communities with their venues and number of articles in Semantic Scholar (S2) / AMiner (AM) along with quantitative clustering evaluation results

	Dataset		Clustering Quality			
Label	Venues	# Articles	Doc2Vec	tSNE	PCA	MODiR
Data Mining	KDD, ICDM, CIKM, WSDM	4,728 / 13,699	0.49 / 0.39	0.30 / 0.55	0.52 / 0.55	0.39 / 0.42
Database	SIGMOD, VLDB, ICDE, EDBT	7,155 / 14,888	0.49 / 0.82	0.64 / 0.34	0.47 / 0.34	0.69 / 0.32
ML	NeurIPS, AAAI, ICML, IJCAI	10,374 / 41,815	0.51 / 0.35	0.21 / 0.23	0.38 / 0.23	0.35 / 0.23
NLP	EMNLP, ACL, CoNLL, COLING	41,815 / 22,523	0.58 / 0.76	0.73 / 0.34	0.81 / 0.34	0.73 / 0.68
Comp Vision	CVPR, ICCV, ICIP, SIGGRAPH	11,898 / 43,558	0.51 / 0.67	0.56 / 0.39	0.49 / 0.39	0.54 / 0.29
HCI	CHI, IUI, UIST, CSCW	8,608 / 33,615	0.64 / 0.68	0.47 / 0.41	0.61 / 0.41	0.39 / 0.38
Average	–	–	0.54 / 0.61	0.49 / 0.37	**0.54** / 0.38	0.53 / **0.39**

Table 4: AtEdge-length of resulting graph layouts

Algorithm	Aminer	SemanticScholar	Enron
tSNE	5.32	4.09	3.89
PCA	5.00	3.91	3.60
MODiR	**4.79**	**2.94**	**2.59**

readability [31]. While these are very elaborate and consider many aspects, we decided to use Noack's normalised AtEdge-length [32]:

$$AtEdge = \frac{\sum_i \sum_j \|\pi_i - \pi_j\|}{|E|} \Big/ \frac{\sum_i \sum_j \|\pi_i - \pi_j\|}{|\mathcal{P}|^2}.$$

It describes how well the space utilisation is by measuring whether edges are as short as possible with respect to the size and density of the graph. Table 4 contains the results.

Although the AtEdge metric is comparable for layouts of the same graph, it is not comparable between datasets as can be seen by the fact, that a larger number of edges causes an overall lower score. The AtEdge length produced by PCA is generally better than that of tSNE while *MODiR* outperforms both as our approach specifically includes an optimised network layout. The better performance of PCA over tSNE can be explained by the resulting layouts being more densely clustered in one spot. Although the AtEdge length aims to give a lower score for too close positioning, it is not able to balance that to the many very long edges in the layout produced by tSNE.

4.4 Qualitative Evaluation

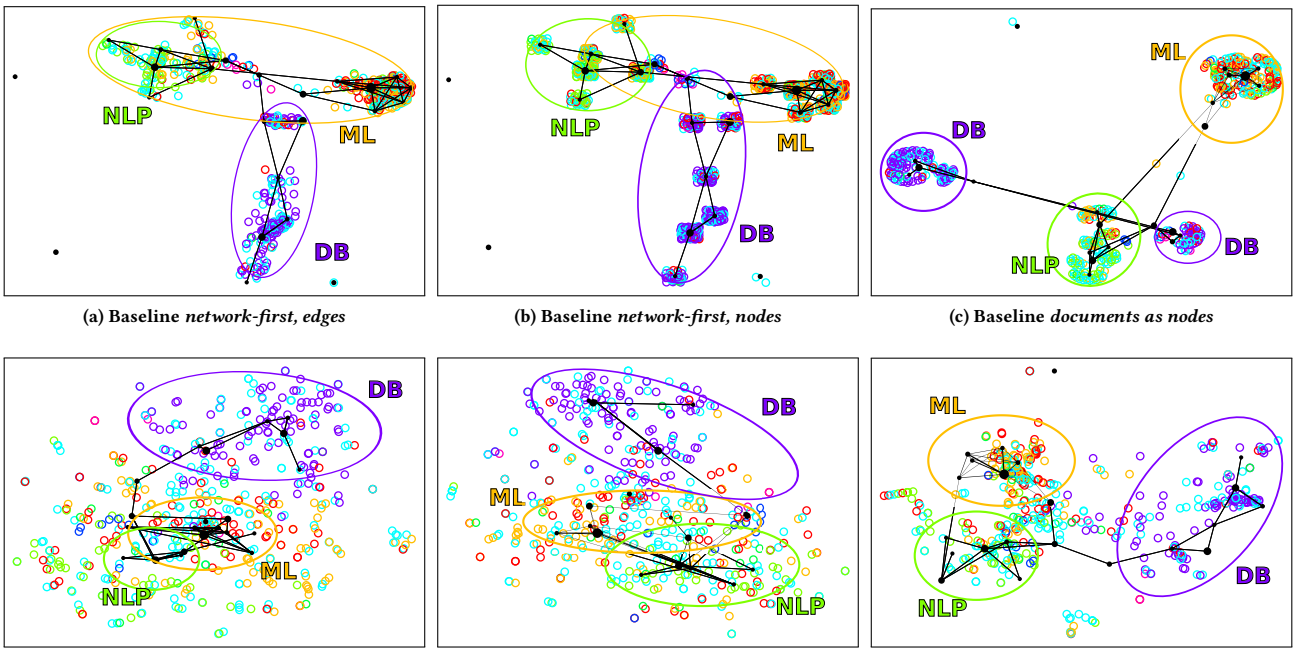
Apart from a purely quantitative evaluation, we use the hand-selected Semantic Scholar dataset (S2b) to visually compare compare network-centric baselines (a-c), document-focused baselines (d-e) and *MODiR* (f) in Figure 3. Papers are depicted as circles where the stroke colour corresponds to the communities, black lines and dots are authors and their co-authorships, size corresponds to the number of publications. For better readability and comparability, the number of drawn points is reduced and three communities are marked.

In Figure 3a we use the weighted co-authorship network drawn using [16] and scatter the papers along their respective edges after the graph is laid out. We see, that active collaboration is easy to identify as densely populated edges and research communities of selected areas are mostly coherent and unconnected researchers are spatially separated from others. Although it is possible to distinguish the different communities in the graph layer,the document landscape is not as clear. The ML researchers are split apart from the rest of the NLP community, which in turn is overcrowded. Figure 3b uses the same network layout but places articles randomly around their first author, which makes it easy to spot the scientific communities by colour. Lastly, we include papers as nodes and co-authorship edges are connected through them during the network layout in Figure 3c. This produces a very clean looking layout compared with the other baselines, however papers lump together and are not evenly distributed. Furthermore, semantic nuances between papers are mostly lost which becomes most apparent in the now separated database clusters. Also, the semantic overlap between the ML and NLP communities is not noticeable.

Figure 3d positions documents using tSNE and places researchers using Equation (1). We see that articles are positioned on the landscape so that research areas are distinctly recognisable by colour. Papers that could not be assigned to a specific area are scattered across the entire landscape. The collaboration network is laid out surprisingly good. The research interests of the authors are coherent between the network and the document landscape, it even shows the close relation between NLP and ML, while showing a clear separation to database related topics. Nonetheless, the network should be loosened for better readability, for example members of the same research group who frequently co-author papers tend to collide.

Unconnected authors are almost not visible as they drift toward densely populated areas in the middle. In Figure 3e, we included authors as virtual documents as the sum of their papers during the tSNE reduction. This shows some improvement, as the network layout is more loose and fewer edges overlap and the issue with collapsing research groups is also mostly mitigated. The semantic overlap of ML and NLP is niceley captured along with the difference to the database papers. However, the network is not clearly readable.

With *MODiR*, the three research communities become clearly distinguishable, both in the graph layer and in the document landscape. Nodes of well connected communities are close together, yet are not too close locally, and separate spatially from other communities. The document landscape is laid out more clearly, as papers from different fields are grouped to mostly distinct clusters. Obviously



(a) Baseline *network-first, edges* (b) Baseline *network-first, nodes* (c) Baseline *documents as nodes*

(d) Baseline *documents-first* (e) Baseline *node+document tSNE* (f) *MODiR*

Figure 3: Semantic Scholar co-authorship network (S2b), subsampled for readability; (a) the network is laid out first, documents are randomly placed along edges; (b) the network is laid out first, documents are randomly placed around nodes; (c) documents are part of the network layout as nodes in the graph that replace author-author edges; (d) the document landscape is laid out first, nodes are positioned at the centre of their associated documents; (e) tSNE is applied on papers and authors together, where documents are aggregated to represent authors

there is still a slight overlap as a result of semantic similarities. As previously pointed out, this visualisation also correctly reveals, that the ML and NLP communities are more closely related to each other (both use machine learning) than to DB. The authorship of documents however can only be conveyed through interaction, so this information is not present in the static visualisations shown here. Based on these results we argue, that the network information improves the (visual) community detection. The document embeddings of articles can only reflect the semantic similarities, which may overlap. In conjunction with information from the co-authorship network, the underlying embeddings are put into their context and thus are more meaningful in a joint visualisation.

Further, we provide additional visualisations of our algorithms with more data. Figure 4 shows the academic corpus (S2), from which we selected all papers from co-authors around high-impact authors from six research communities as described above. We see how the network information influences the landscape and communities become clearly visible. Although the global structure of both semantics and network is readable, an additional objective to discourage overlapping edges could further improve the result. For better interpretability we used a baseline approach to extract position-based keyphrases to overlay them on the landscape. Our prototype offers the user very basic interactions to explore the landscape, such as zooming, panning, highlighting parts of the landscape where a search term appears, or looking up entities and categories (if available).

5 CONCLUSIONS AND FUTURE WORK

In this paper we discussed how to visualise large document collections by jointly visualising text and network aspects on a single canvas. To this end, we identified three principles that should be balanced by a visualisation algorithm. From those we derived formal objectives that are used by a gradient descend algorithm. We have shown how to use that to generate landscapes which consist of a base-layer, where the embedded unstructured texts are positioned such that their closeness in the *document landscape* reflects semantic similarity. Secondly, the landscape consists of a *graph layer* onto which the inherent network is drawn such that well connected nodes are close to one another. Lastly, both aspects can be balanced so that nodes are close to the documents they are associated with while preserving the graph-induced neighbourhood. We proposed *MODiR*, a novel multi-objective dimensionality reduction algorithm which iteratively optimises the document and network layout to generate insightful visualisations using the objectives mentioned above. In comparison with baseline approaches, this multi-objective approach provided best balanced overall results as measured by various metrics. In particular, we have shown that *MODiR* outperforms state-of-the-art algorithms, such as tSNE. We

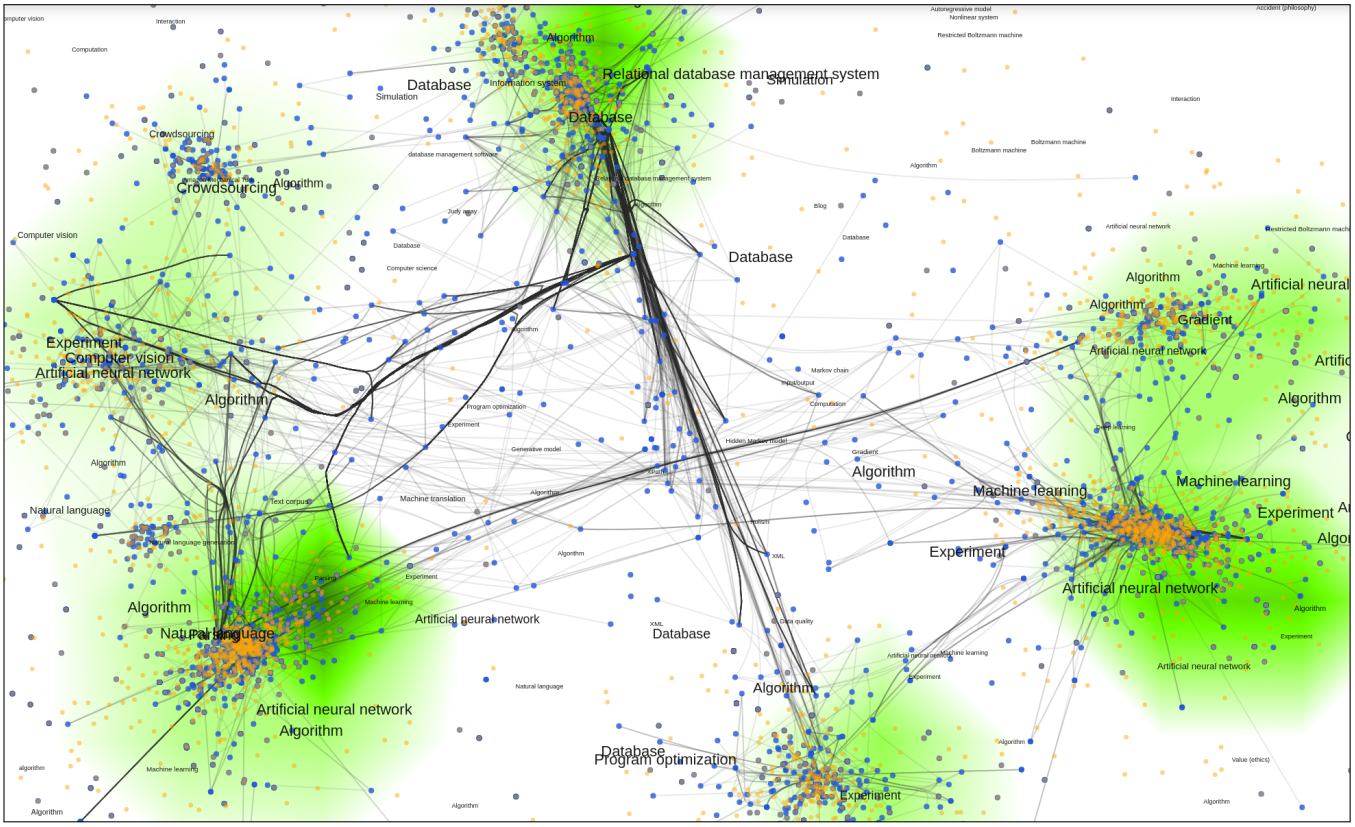


Figure 4: *MODiR* visualisation of Semantic Scholar (S2), all six communities become clear. Authors are blue dots, papers are orange dots, green density map is based on all papers, black opaque edges connect co-authors.

also implemented an initial prototype for an intuitive and interactive exploration of multiple datasets as shown in Figure 4 We have shown the effectiveness of *MODiR* using a number of different large document collections by measuring the topical clustering quality of the document landscape and the network layout of the graph layer. Additionally we used different visualisations to inspect calculated layouts.

While our prototype of *MODiR* allows basic interactions, we look into improving the look-and-feel further in future work. For easy interpretability and fast exploration we found it useful to have an overlay of keywords. These help to semantically distinguish different areas of the landscape. In our preliminary work we used tf-idf on meta-documents, for which we concatenate actual documents. The simplest approach aggregates documents within a cell of a virtual grid across the landscape. Our more advanced approach, as used in the example shown above, uses density based clustering to group documents. Furthermore, we used established keyphrase extraction algorithms instead of selecting words with the highest tf-idf score. However, in all our experiments we see room for improvement as words seem repetitive or not relevant enough. In future work we hope to focus on the problem of selection and placement of descriptive keywords or keyphrases. This will improve the way users are able to navigate the landscape.

REFERENCES

[1] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in Semantic Scholar. In *NAACL-HLT*. MIT Press, Cambridge, MA, USA, 84–91.

[2] David Arthur and Sergei Vassilvitskii. 2007. k-means++: The advantages of careful seeding. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM Publications, Philadelphia, USA, 1027–1035.

[3] Dmitry Baranchuk, Artem Babenko, and Yury Malkov. 2018. Revisiting the inverted indices for billion-scale approximate nearest neighbors. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer-Verlag, Heidelberg, Germany, 202–216.

[4] Sandro Cavallari, Vincent W Zheng, Hongyun Cai, Kevin Chen-Chuan Chang, and Erik Cambria. 2017. Learning community embedding with community detection and node embedding on graphs. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*. ACM Press, Geneva, Switzerland, 377–386.

[5] Marie-Anne Chabin. 2017. Panama papers: a case study for records management? *Brazilian Journal of Information Science: Research Trends* 11, 4 (2017), 10–13.

[6] Shiyu Chang, Wei Han, Jiliang Tang, Guo-Jun Qi, Charu C Aggarwal, and Thomas S Huang. 2015. Heterogeneous network embedding via deep architectures. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. ACM Press, Geneva, Switzerland, 119–128.

[7] 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.

[8] Mark Coddington. 2015. Clarifying journalism's quantitative turn: A typology for evaluating data journalism, computational journalism, and computer-assisted reporting. *Digital Journalism* 3, 3 (2015), 331–348.

[9] Tommy Dang and Angus Forbes. 2017. CactusTree: A tree drawing approach for hierarchical edge bundling. In *Proceedings of the IEEE Pacific Visualization Conference (PacVis)*. IEEE, New York City, USA, 210–214.

[10] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using Structured Events to Predict Stock Price Movement: An Empirical Investigation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. MIT Press, Cambridge, MA, USA, 1415–1425.

[11] Alon Efrat, Yifan Hu, Stephen G Kobourov, and Sergey Pupyrev. 2015. MapSets: Visualizing embedded and clustered graphs. *JGAA* 19, 2 (2015), 571–593.

[12] Evgeniy Faerman, Felix Borutta, Kimon Fountoulakis, and Michael W Mahoney. 2018. LASAGNE: Locality And Structure Aware Graph Node Embedding. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI) (WIC)*. IEEE, New York City, USA, 246–253.

[13] Blaz Fortuna, Marko Grobelnik, and Dunja Mladenic. 2005. Visualization of text document corpus. *Informatica* 29, 4 (2005), 497–502.

[14] Katrin Franke and Sargur N Srihari. 2007. Computational forensics: Towards hybrid-intelligent crime investigation. In *Proceedings of the International Symposium on Information Assurance and Security (IAS)*. IEEE, New York City, USA, 383–386.

[15] Daniel Fried and Stephen G Kobourov. 2014. Maps of computer science. In *Proceedings of the IEEE Pacific Visualization Conference (PacVis)*. IEEE, New York City, USA, 113–120.

[16] Thomas MJ Fruchterman and Edward M Reingold. 1991. Graph drawing by force-directed placement. *Software: Practice and experience* 21, 11 (1991), 1129–1164.

[17] Martin Gronemann and Michael Jünger. 2012. Drawing clustered graphs as topographic maps. In *Proceedings of the International Symposium on Graph Drawing (GD)*. Springer-Verlag, Heidelberg, Germany, 426–438.

[18] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Representation Learning on Graphs: Methods and Applications. *IEEE Data Engineering Bulletin (DEB)* 40, 3 (2017), 52–74.

[19] Jan Hildenbrand, Arlind Nocaj, and Ulrik Brandes. 2016. Flexible level-of-detail rendering for large graphs. In *GD*. Springer-Verlag, Heidelberg, Germany, 625–627.

[20] Geoffrey E Hinton and Sam T Roweis. 2003. Stochastic neighbor embedding. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*. NIPS Foundation, Inc., San Diego, USA, 857–864.

[21] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLOS ONE* 9, 6 (2014), e98679.

[22] Mukundan Karthik, Mariappan Marikkannan, and Arputharaj Kannan. 2008. An intelligent system for semantic information retrieval information from textual web documents. In *International Workshop on Computational Forensics (IWCF)*. Springer-Verlag, Heidelberg, Germany, 135–146.

[23] Bryan Klimt and Yiming Yang. 2004. The Enron corpus: A new dataset for email classification research. In *Proceedings of the European Conference on Machine Learning (ECML)*. Springer-Verlag, Heidelberg, Germany, 217–226.

[24] 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.

[25] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning (ICML)*. JMLR Inc. and Microtome Publishing, Brookline, USA, 1188–1196.

[26] 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.

[27] Jie Liu, Zhicheng He, Lai Wei, and Yalou Huang. 2018. Content to node: Self-translation network embedding. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. ACM Press, Geneva, Switzerland, 1794–1802.

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

[29] Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. (2018), 44 pages.

[30] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*. NIPS Foundation, Inc., San Diego, USA, 3111–3119.

[31] Quan Hoang Nguyen, Peter Eades, and Seok-Hee Hong. 2017. Towards faithful graph visualizations. (2017).

[32] Andreas Noack. 2007. *Unified quality measures for clusterings, layouts, and orderings of graphs, and their application as software design criteria.* Ph.D. Dissertation. Brandenburg University of Technology, Cottbus-Senftenberg, Germany. Chapter 6, pp97.

[33] Patrick Cheong-Iao Pang, Robert P Biuk-Aghai, Muye Yang, and Bin Pang. 2017. Creating realistic map-like visualisations: Results from user studies. *JVLC* 43 (2017), 60–70.

[34] Karl Pearson. 1901. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* 2, 11 (1901), 559–572.

[35] 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.

[36] Helen C Purchase. 2002. Metrics for graph drawing aesthetics. *JVLC* 13, 5 (2002), 501–516.

[37] Tim Repke and Ralf Krestel. 2018. Bringing back structure to free text email conversations with recurrent neural networks. In *Proceedings of the European Conference on Information Retrieval (ECIR)*. Springer-Verlag, Heidelberg, Germany, 114–126.

[38] Arnaud Sallaberry, Yang-chih Fu, Hwai-Chung Ho, and Kwan-Liu Ma. 2016. Contact trees: Network visualization beyond nodes and edges. *PloS one* 11, 1 (2016), e0146368.

[39] John W Sammon. 1969. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* 100, 5 (1969), 401–409.

[40] Jörg Schlötterer, Christin Seifers, and Michael Granitzer. 2017. On Joint Representation Learning of Network Structure and DOcument Content. In *International IFIP Cross Domain Conference for Machine Learning & Knowledge Extraction (CD-MAKE)*. Springer-Verlag, Heidelberg, Germany, 237–251.

[41] Shilad Sen, Anja Beth Swoap, Qisheng Li, Brooke Boatman, Ilse Dippenaar, Rebecca Gold, Monica Ngo, Sarah Pujol, Bret Jackson, and Brent Hecht. 2017. Cartograph: Unlocking spatial visualization through semantic enhancement. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*. ACM Press, Geneva, Switzerland, 179–190.

[42] Dominic Seyler, Tatiana Dembelova, Luciano Del Corro, Johannes Hoffart, and Gerhard Weikum. 2018. A Study of the Importance of External Knowledge in the Named Entity Recognition Task. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. MIT Press, Cambridge, MA, USA, 241–246.

[43] 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.

[44] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale information network embedding. In *Proceedings of the International World Wide Web Conference (WWW)*. ACM Press, Geneva, Switzerland, 1067–1077.

[45] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. ACM Press, Geneva, Switzerland, 990–998.

[46] Nees Jan Van Eck and Ludo Waltman. 2014. Visualizing bibliometric networks. In *Measuring scholarly impact*. Springer-Verlag, Heidelberg, Germany, 285–320.

[47] Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. 2010. TIARA: a visual exploratory text analytic system. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. ACM Press, Geneva, Switzerland, 153–162.

[48] Jaewon Yang and Jure Leskovec. 2014. Overlapping communities explain core–periphery publisher of networks. *Proc. IEEE* 102, 12 (2014), 1892–1902.

[49] Daokun Zhang, Jie Yin, Xingquan Zhu, and Chengqi Zhang. 2018. Network Representation Learning: A Survey. *IEEE Transactions on Big Data (TBD)* Early Access (2018), 1–25.