CGV: Medium: Collaborative Research: SMAP—Topic Maps with Guarantees

Katy Börner and Robert Light (Indiana University) Stephen G. Kobourov (University of Arizona)

1 Introduction

Today, anyone on the Internet has access to much of humankind's collective knowledge and expertise. While this grand flow of information provides a plethora of opportunities, it can also be an overwhelming deluge making it urgent that we answer the question of how best to amplify human perception and cognition to navigate, manage, and make sense of this ever-expanding, advancing knowledge. For decades, topic maps have been employed as visual interfaces to textual data, complete with data overlays, *e.g.*, expertise profiles of researchers, institutions, or entire countries. Topic maps showing all sciences, also called science maps, are used by a wide range of professionals to grasp crucial developments in science and technology; see Fig. 1. Interestingly, though, little is known about the readability, accuracy, and utility of topic maps or science maps. Many different generation algorithms and techniques exist, see review of three widely used topic map generation workflows in Section 2. However, there is no consensus as to which one works best, or what the word "best" truly means in this context.

This project aims to provide answers to these questions. In *Phase 0* we address the first and most basic task – modularizing existing tools and software frameworks, so well-defined, comparable workflows can be composed for the many common tasks (e.g., text preprocessing, dimensionality reduction, clustering). In *Phase 1* we analyze the best text preprocessing techniques, dimensionality reduction techniques, clustering, and removal of text overlaps. Also in this phase, the resulting clusters and 2D placements will be subjected to validation. Quantitative validation will be obtained through measuring the embedding (stress, distortion), the layout (compactness and distribution), and the clustering (accuracy and cohesion). Qualitative validation will be obtained through human subject experiments to assessing the legibility and utility of the topic spaces. In *Phase 2* we will validate the different visualization techniques (point clouds, node-link network diagrams, map-like spatializations) and different data overlays (cartogram, choropleth, heatmap). Quantitative analysis will ensure that certain guarantees are maintained, while user-based qualitative studies will answer questions about engagement, knowledge retention, and ease of use.

The ultimate goal of this project is to create qualitatively and quantitatively validated topic maps, based on validated data sets, and made available to the general public via a state-of-the-art online visualization system. Creating accurate and easy to understand maps would make it easier to communicate complex topical concepts to researchers, students, grant and research agencies, industry, data providers, and the general public as a whole.

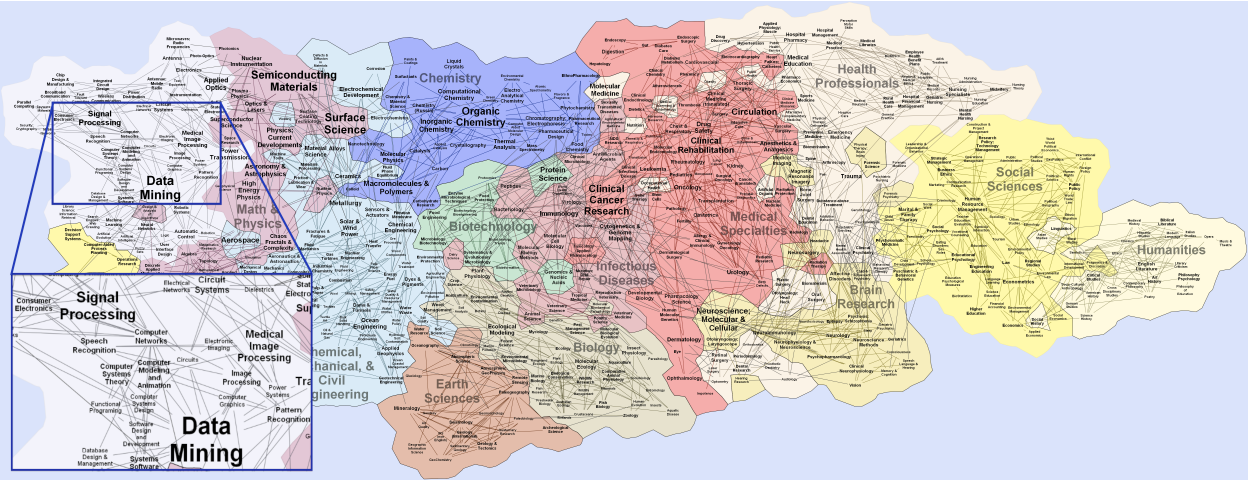

Figure 1: The UCSD map of science visualized with GMap.

2 Prior Work: Topic Map Generation and Validation

Rendering imagined and topical spaces as a map has a long history, *e.g.*, the 1930s Map of Middle Earth by Tolkien [131]. The Atlas of Science features a 22-page timeline showing the evolution of science maps—from the first hand-rendered maps in the 1930s to multi-layer, interactive visualizations in 2007 [17]. A review of different topic analysis and visualization techniques can be found in [19] and [111].

We review the three commonly used textual analysis and visualization workflows. They were developed by cartographers, scientometricians, and graph theorists. As we are interested in comparing these three approaches we first review them and then attempt a unification of terminology and an alignment of key workflow steps.

2.1 Self-Organizing Maps (SOMs) and Geographic Information Systems (GIS)

A Self-Organizing Map (SOM), developed by Kohonen [104], is a type of neural network, trained using unsupervised learning to produce 2D representation of the input space of the training samples. A SOM produces a 2D surface tiled with square or hexagonal cells and is, in a sense, an alternative approach to traditional dimensionality reduction techniques such as MDS [105] and PCA [97]. The input to a SOM is a collection of n-dimensional vectors $E = \{e_1, \ldots, e_m\}$ and each cell in the map is associated with an n-dimensional vector $\mathbf{t}_j$. The smallest Euclidean distance between an input vector and a cell determines the placement of the input: $dist(\mathbf{e}_i, \mathbf{t}_j) = \sqrt{\sum_{k=1}^{n} (\mathbf{e}_{ik} - \mathbf{t}_{jk})^2}$. Initially cell vectors are random but during the training of the network they change in order to better fit the input values. The training consists of repeatedly inserting input values into the network and updating the corresponding cell and its neighbors: $\mathbf{t} \leftarrow \mathbf{t} + \alpha \cdot \beta(\mathbf{t} - e)$. The learning coefficient α is used to allow greater changes to the cell value in early stages of the computation, and smaller changes in later stages. The neighbor coefficient β is used to decrease the effect of the learning for distant cells and also changes over time, so that the radius of affected neighbors decreases in later stages. In this manner, the input data is repeatedly inserted in the SOM, on the order of millions of times. Once the network is trained, each element from the input is inserted in the best-fit cell. Note that the number of cells in the map is fixed in advance and it is possible that many cells are empty, or contain hundreds or thousands of elements.

Skupin[125] developed a general workflow (Fig. 2) that uses self-organizing maps (SOM) and geographic information systems (GIS) software to render two-dimensional maps from textual documents. (Key)words from the documents are used as terms in the creation of a term-document matrix. For each

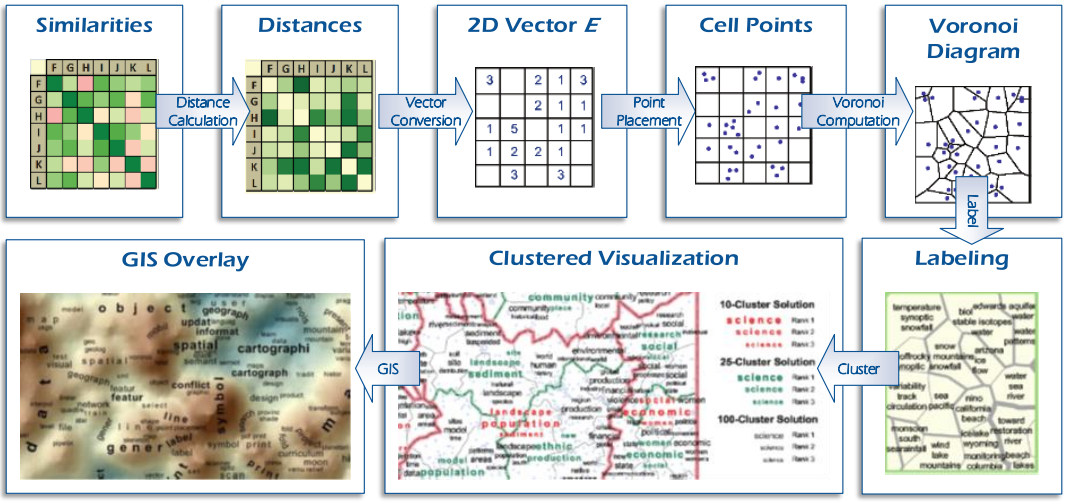

Figure 2: Self-Organizing Map (SOM) and Geographic Information System (GIS) workflow

document its corresponding row vector in the matrix is used to train the SOM. After several hours of training, the papers are inserted into the map. The papers within a particular cell are randomly placed and a Voronoi diagram of the cell is computed. A hierarchical clustering merges together similar cells. Labels are computed based on term frequency. GIS software is used to render the final representation.

As pointed out by the author, better stemming to avoid keyword duplication, better ways to deal with sparse vectors, and better clustering to prevent misleading result, are needed. In later work, Skupin and de Jongh [128] consider a different set of papers using document titles rather than keywords as terms and counting term occurrence in the full texts, rather than just the abstracts. The training for 6 million iteration for their 50×50 SOM required over five days. Clustering data in SOM representations of research papers has posed problems in the early approaches. Skupin [126] attempts to address some of these problems with different clustering methods, such as k-means, hierarchical clustering, and neuron label clustering. As pointed out by the author, there is no single best approach, but that different ones have different advantages and disadvantages. A very recent paper by Skupin *et al.* [127] applies the SOM approach to a much larger dataset with over 2 million publications, each of which is tagged with at least five PubMed Medical Subject Headings (MeSH). The final SOM has 275×275 cells, only 10% of the vocabulary is used, and the vectors associated with each paper are binary (rather than reflecting term frequencies). Still the computation would have required over year, but was accomplished in more reasonable time using a super-computer.

2.2 Graph-Maps (GMap)

The GMap approach combines graph layout and graph clustering, together with appropriate coloring of the clusters and creating boundaries based on clusters and connectivity in the original graph and can be applied to any relational data set [91].

The *GMap framework*, from similarity matrix to the final map, is summarized in Fig. 3. Since the overall input to the GMap framework algorithms is a relational data set, a graph $G = (V, E)$ is extracted from the distance matrix. The vertex set V corresponds to data objects (e.g., research papers) and the edge set E corresponds to relationships between pairs of objects (e.g., topical similarity between two research papers). There are four distinct phases in the GMap framework:

1. Obtain an embedding of the graph in the plane using a multi-dimensional embedding algorithm, *e.g.* MDS, for large graphs [81] to get an initial layout, and then remove overlaps to get the final layout;
2. Perform a cluster analysis to group vertices into clusters, for example using modularity clustering [121] or geometric clustering [112], which will be used to distinguish countries in the map;

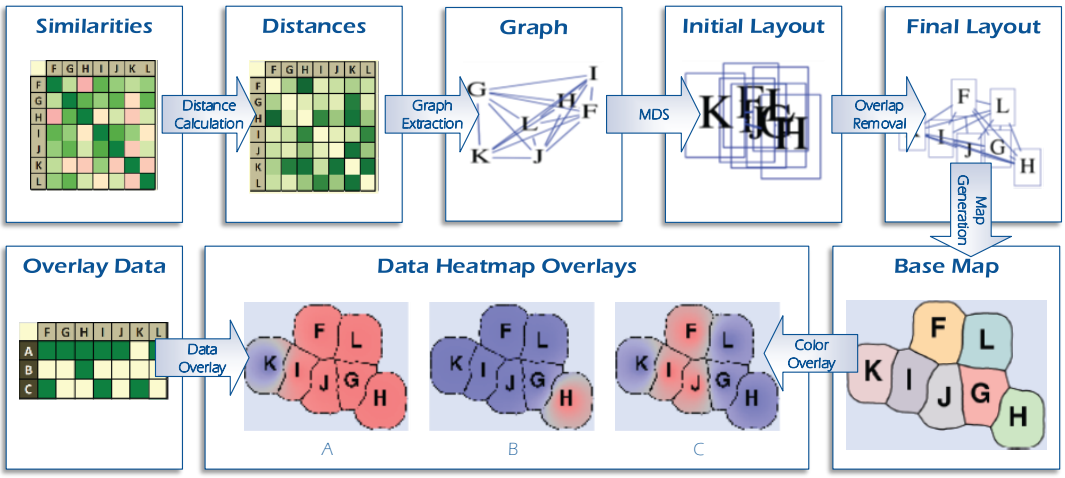

Figure 3: Graph-Map (GMap) workflow

3. Create a geographic map corresponding to the data set, based on a modified Voronoi diagram of the vertices, which in turn is determined by the embedding and clustering (countries are created from clusters, and continents and islands are created from groups of neighboring countries);
4. Assign unique colors to each country, ensuring that neighboring countries have different colors.

The GMap framework is used to create maps of computer science (MoCS) [80]. These maps are generated from words and phrases extracted from the titles of the 2,184,720 research papers in the DBLP bibliography server [108]. The process of extracting topics from the titles relies on basic natural language processing techniques and similarities between them are computed based on co-occurrence. In particular, words from the titles are used as terms in the creation of a term-document matrix. The set of top terms are ordered by importance and the top terms are selected for including in the map representation. Once the set of top terms is selected, pairwise similarity values between top terms are calculated, so that terms that refer to the same or similar topic, or topics that are closely associated, receive high similarity values. Terms are compared using cosine similarity of the feature vectors to produce a matrix of pairwise similarities between terms.

2.3 Topic Modeling and Force-Directed Layout (TMFDL)

The general workflow for topic modeling plus force-directed layout (TMFDL) is shown in Fig. 4. There are many ways to go about creating topics from a body of text or collection of textual documents. Previous projects undertaken by the Börner team have implementing Latent Dirichlet Allocation (LDA) [13]. Rather than trying to assign documents to bins one by one, LDA assumes that all documents are formed from a combination of topics. By iterating through a number of cycles, each time using the previous result as prior knowledge, the algorithm refines the topics to improve the quality of the document-topic fit.

Multiple software packages exist that implement LDA. MALLET, a package developed at the University of Massachusetts, has been utilized by the team. MALLET is an open-source Java-based package for natural language processing and topic modeling [116]. This package allows for document classification, sequence tagging and topic modeling using a variety of algorithms, including LDA as well as hierarchical LDA.

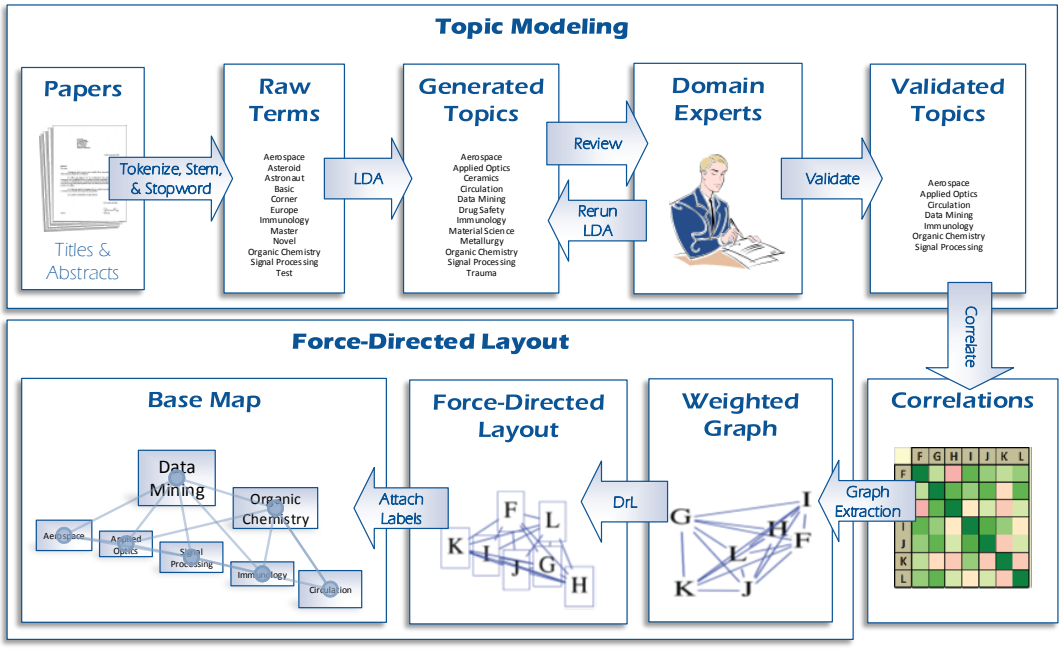

Figure 4: Topic Modeling and Force-Directed Layout workflow

Datasets were processed into a format with an identifier attached to a single block of text containing the title and abstract. This data was stemmed, tokenized and stopworded and input into the LDA algorithm. Initial sets with a pre-determined number of topics were generated. These topics were manually reviewed and domain-specific stopwords identified. These were identified by domain experts and removed before the topics regenerated.

These topics were then reviewed and given names manually based on primary keywords and the most strongly associated data points. These topics then underwent another round of review to determine whether the topic set was too broad or too narrowly defined and the process was repeated as needed with a new number of topics. Once domain-specific experts reached agreement that the topic set was valid, a network was constructed using the topics as nodes and the correlation between those topics with regards to the description of data points as edges. This was a strongly connected graph by default as all data points were automatically mapped to all topics, so the weakest edges were pruned to give the graph definition.

While both SOM and GMap based approaches include techniques to reduce dimensionality from a topical space into a two-dimensional space, LDA does not. The TMFDL approach uses a force-directed layout to distribute the topics across a two-dimensional space as nodes. Edges were created between nodes with weights based on the correlation of those topics over the entire corpus of text. These edges were then treated as forces (or springs, if you wish) trying to pull the nodes in various directions. Edges with greater weights were equated to stronger, stiffer springs that pulled harder at the nodes on either end. The algorithm then released the nodes and let the springs pull them into new positions, searching for a solution that put minimal stress on the springs as a whole. This generally minimized the number of long or crossing edges, creating an aesthetically pleasing layout.

The software that was used to generate these layouts was VxOrd, originally developed at Sandia National Laboratories. This software implemented an algorithm known as DrL. DrL is one of the few force-directed algorithms that can scale to the orders of magnitude required to process the scale of datasets that these projects require, although small-scale graphs did sometimes suffer aesthetically. ISI journal classification is used to compute the clusters [30].

2.4 Validation Studies

Prior work by Börner and colleagues validated different similarity measures, compared text-based vs. linkage based approaches, and aimed to understand the legibility of different maps.

In 2007, a study was performed to map the structure of the fields of Chemistry as well as their change over time [29]. This study is built on a base map of science generated in 2002 using 1.07 million papers across 7,227 journals from the Science Citations Index Expanded (SCIE) and the Social Science Citation Index (SSCI). Bibliographic coupling was generated at the paper level then aggregated by journal. The 7,227 journals were placed on a map using the VxOrd graph layout algorithm. These were then divided into 671 clusters. The coupling counts were further aggregated over the clusters and then returned to the VxOrd algorithm to produce the final layout. The effort to map the evolution of Chemistry over time introduced several challenges, most notably what do to with the 2,350 relevant journals that did not exist in 2002. The decision was made to hold the map static over time and assign the journals to existing clusters based on their coupling to journals already assigned to those clusters. Challenges also emerged when trying to visualize so many clusters of journals, leading to manual aggregation of clusters and reliance on outside data (JCR) in establishing what qualified as a chemistry journal. In spite of these difficulties, this study succeeded in reflecting broad trends regarding the quickly growing impact of BioChemistry and BioEngineering over the prior thirty years, while the more established field of Chemistry showed slower growth.

In 2011, a corpus of metadata from over two million MEDLINE documents (data is available at `http://sci.cns.iu.edu/sts`) was studied to assess the effectiveness of text-based similarity methods. Five different methods were applied over either MeSH headings, titles and abstracts or both to create nine different clustering solutions [129]. Findings showed that the BM25 approach, when applied to the titles and

abstracts, significantly outperformed the more commonly used tf-idf over the same word-document matrix. PubMed's inherent ranking algorithm (PMRA) showed the best performance in terms of coverage and coherence, but was specially designed to optimally use PubMed data. Topic modeling also outperformed tf-idf over titles and abstracts, but did not perform well for smaller "fine grain" clusters in a large dataset. While MeSH based algorithms were less computationally intensive, but generally showed less accuracy or coherence than techniques that employed the title and abstract. Only tf-idf showed largely identical performance for either type of input.

In 2012, Börner, Light and colleagues computed the most comprehensive and up to date science map available today [22]. This work built on the UCSD base map that had been created from about 16,000 source journals in 2005. This map was generated using bibliographic coupling of both references and shared keywords at the paper level and was clustered at multiple levels, creating 554 journal-defined subdisciplines aggregated into 13 disciplines. The updated UCSD science map and classification system covers 10 years (2001-2010) of Web of Science data and 8 years (2001-2008) of Scopus data with subdiscipline assignments by SciTech Strategies. The map was computed by measuring the bibliographic coupling from each journal to the existing subdisciplines, corrected for the volume of publications within each discipline (to keep larger subdisciplines from overwhelming similar but smaller counterparts). Efforts were made to simplify the map by reducing the number of journals that were fractionally assigned to multiple subdisciplines. The added journals increase the influence of the social sciences and humanities on the map from 19 to 35%. All data associated with the map has been openly published at `http://sci.cns.iu.edu/ucsdmap`, allowing researchers to utilize the base map in visualizing their own data. This map will be used in Phase 2 of the proposed work.

3 Planned Work—Phase 0: Algorithm Modularization

In order to compare different algorithms, the entire analysis pipeline has to become modular. In other words, each software system has to be broken into its parts and wrapped in such a way that we can interchange them quickly and easily. Börner's team has been working on plug-and-play macroscope tools [18] for more than 10 years. The OSGi industry standard, in combination with the CIShell framework, provide an easy and scalable means to wrap code into plug-ins and to deploy them in tools. Building all of the functions into Sci2 plugins not only allows for documented workflows, but will also builds them into a free and open source system that can be easily shared for research and learning purposes. All code developed in this project will be incorporated in the Sci2 suite of macroscopes (`http://cishell.org`) extending the functionality of this widely used tool considerably.

4 Planned Work—Phase 1: Topic Analysis

As discussed in Section 1, Phase 1 reads different datasets, calculates similarities, and reduces dimensionality of text into a meaningful topic space with validation (Fig. 5). Each step will be detailed subsequently.

4.1 Textual Datasets

Phase 1 will use four datasets, the first three of which come from the Scholarly Database [107] a repository of free government information maintained by Indiana University:

1. *MEDLINE papers* set was generated as a combination of MEDLINE and Scopus (Elsevier) data. Records were limited to documents published between 2004 and 2008 that had an abstract that contained at least 5 MeSH terms with a bibliography of at least 5 references. This resulted in a corpus of 2,153,769 unique documents. Information was gathered including title, abstract, MeSH terms, and reference lists. Data is freely available at `http://sci.cns.iu.edu/sts`.
2. *NSF grant dataset* includes 374,467 grants. Since 1988, over 90% of these contain an abstract, with 55,616 from 2004 to 2008. These grants are linked to 300,168 articles via the NSF database.

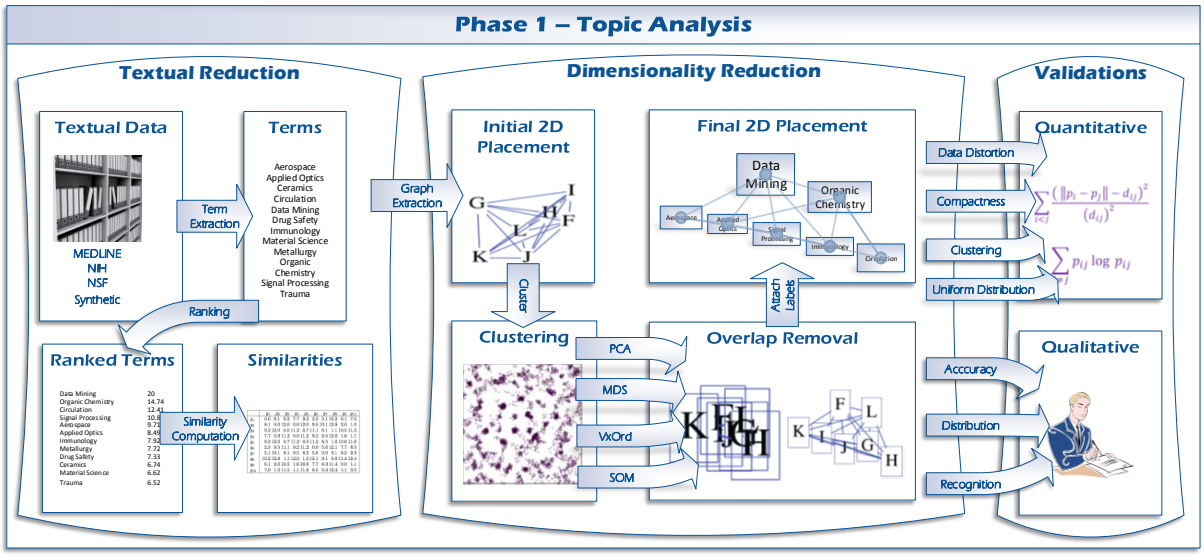

Figure 5: Phase 1—Topic Analysis

3. *NIH grant dataset* includes 2,490,837 grants dating from 1972 to 2012. Of these, 78 percent contain an abstract. These grants are linked to 1,378,728 MEDLINE articles via the NIH database. There are 118,782 grants with abstracts from 2004 to 2008 that is covered by the MEDLINE dataset.
4. *Synthetic data* will be generated with certain preconditions in order to test the map production pipelines ability to capture the inherent qualities of a dataset. Varying clusters will be generated systematically in order to ensure that algorithms are correctly identifying known entities within a dataset.

4.2 Textual Reduction

In the next step we process the text data, in the form of titles and abstracts of research papers or research grants, in order to find related texts. In the following discussion we use just the titles but in general, the described methods will also be used for the richer datasets where we also have abstracts.

Term Extraction: Multi-word terms need to be extracted from the titles of the papers. Part of speech (POS) tags are used to choose words that constitute topically meaningful terms, and exclude functional words (words conveying little semantic meaning, such as "the" and "and"). The Natural Language Toolkit (NLTK) POS tagger [11] can be used to label the words in all titles with POS tags. Once a title is tagged, maximal subsequences of words with POS tags matching the following regular expression can be extracted from titles: $(\langle JJ\rangle|\langle JJR\rangle|\langle JJS\rangle|\langle NN\rangle|\langle NNS\rangle|\langle NNP\rangle|\langle NNPS\rangle)$ where JJ, JJR, and JJS are tags representing normal adjectives, comparative adjectives, and superlative adjectives, respectively, while NN, NNS, NNP, and $NNPS$ are nouns, plural nouns, proper nouns, and proper plural nouns, respectively.

Ranking: Each ranking function orders terms by their assigned weight (rank), and the top n of them are selected, where n is the number of terms that will be visualized. We will implement and compare several such ranking functions beyond the basic Term Frequency (tf). Even after removing common stop-words, tf tends to rank highly many semantically meaningless words. Term Frequency-Inverse Document Frequency (tf/idf) addresses this problem by normalizing the frequency of a word by its frequency in a larger text collection. In this domain, terms only occur once per document, so the inverse document frequency of a term is almost always 1. Tf/idf can be modified for such datasets, by treating the entire collection of titles as a single document, and using the term's frequency in a reference corpus from a different domain (e.g., the Brown English Language corpus) for the inverse weighting value. The resulting method is Term Frequency–Inverse Comparison Frequency (tf/icf). *C-value* [79] is another ranking method designed to

account for possible nesting of multi-word terms (where short terms appear as word subsequences of longer terms), by incorporating total frequency of occurrence, frequency of occurrences of the term within other longer terms, the number of types of these longer terms, and the number of words in the term.

Similarity Computation: Once a set of top terms is selected, pairwise similarity values $S(t_i, t_j)$ between top terms t_i and t_j are calculated. We seek similarity functions that measure how closely the topics represented by two terms are related. We use term-document co-occurrence as the basis of these similarity values, assuming that terms that appear together in multiple titles are more likely to be related in meaning. The similarity functions take a term-document matrix, M, as input. Columns of correspond to titles of papers. Rows correspond to terms extracted by the term-extraction step. Each entry is the frequency of occurrences of the term indexed by the entry's row in the title indexed by the entry's column. We will implement and evaluate several similarity functions. *Latent Semantic Analysis (LSA)* [48] is a method of extracting underlying semantic representation from the term-document matrix, M. The singular value decomposition of M is calculated using sparse-matrix methods, and rows in this decomposition represent terms as feature vectors in the high-dimensional semantic space. Terms are compared using cosine similarity of the feature vectors to produce a matrix of pairwise similarities between terms. The *Jaccard coefficient* [96] variants offer alternative similarity functions, to accommodate the nearly boolean nature of the term-document matrix. The Jaccard coefficient calculates pairwise term similarity as the number of documents two terms appeared together in, divided by the number of documents either term appeared in: $Jacc(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}$, where S_i and S_j are the sets of documents that the two terms being compared appeared in.

4.3 Dimensionality Reduction

As in many other problems, the term-similarity matrix is high-dimensional but we would like to display the data in 2D space. As usual, the hope is that high-dimensional data are multiple, indirect measurements of an underlying lower-dimensional distance function. As our ultimate visualization goals are 2D maps of data, we need to reduce the dimensionality of our data. With this in mind, we will compare several different dimensionality reduction techniques, starting with the techniques used in the three earlier map-based visualization frameworks (SOM, GMap, and TMFDL), described in Section 2.

Dimensionality reduction can be thought of as the process of deriving a set of degrees of freedom which can be used to reproduce most of the variability of a data set. Here we assume that we need to reduce our n-dimensional similarity matrix down to 2D space.

Principal components analysis [97] is a classical method that provides a sequence of best linear approximations to a given high-dimensional observation. It is one of the most popular techniques for dimensionality reduction. Given input such as our n-dimensional similarity matrix, PCA aims to find a 2D linear subspace, such that the data points lie on this linear subspace. Such a reduced subspace attempts to maintain most of the variability of the data. The linear subspace is specified by 2 orthogonal vectors that form a new coordinate system, called the "principal components", onto which the variance retained under projection is maximal. The effectiveness of PCA is limited by its global linearity.

An alternative perspective on dimensionality reduction is offered by multi-dimensional scaling [105]. MDS is another classical approach that maps the original high dimensional space to a lower dimensional space, but this time attempting to preserve pairwise distances. The aim is to construct a configuration of n points in 2D Euclidean space such that the pairwise distance between any pair is very similar to the pairwise distance between the two corresponding vectors in the input similarity matrix. Although the mathematics behind MDS is different from that for PCA, it also suffers from the same drawback (global linearity).

Self-organizing maps [104] can be used to perform dimensionality reduction in a non-linear fashion, by learning in an unsupervised way, a projection from the high dimensional space to 2D. Unlike PCA and MDS, SOM does not have an explicit function that is being optimized (recall that PCA optimizes variance and MDS optimizes distances). Moreover, discussed in the prior work section, SOMs are highly sensitive to the two parameters α and β, and there is no way to ensure that the iterative calculation converges.

Force-directed methods [99] can also be used to perform non-linear dimensionality reduction. Also known as spring embedders, such methods calculate 2D positions for input points by treating the input similarity as a weighted adjacency matrix of a graph. This is accomplished with the help of a physical model based on spring forces, similar to those in Hooke's law. For example, there could be repulsive forces between all nodes and attractive forces between nodes which are adjacent. Graphs drawn with these algorithms tend to be aesthetically pleasing. However, similar to the SOM methods, force-directed methods are not guaranteed to converge and their utility is diminished when the underlying data is very large. Recently there have been several new and more scalable methods, including some by the PIs [81, 82].

In our experiments we will use all standard dimensionality reduction techniques, starting with the ones that are part of the three existing frameworks (SOM, GMap, and TMFDL).

4.4 Clustering

In some of our datasets the clustering is already given. For example, the updated UCSD science map and classification system covers 10 years (2001-2010) of Web of Science data and 8 years (2001-2008) of Scopus data, where the clusters are computed based on subdiscipline assignments (already computed by SciTech Strategies).

Even when the data does not come with implicit clustering information, all three of the existing visualization frameworks use some implicit or explicit clustering scheme. The SOM-based methods implicitly create neuron-label clusters, or explicitly create clusters using k-means [112] or hierarchical clustering methods. GMap-based methods rely of graph clustering techniques such as modularity-clustering [121], which produces good results when paired up with force-directed or MDS-type embedding [122]. The TMFDL-based method uses ISI journal classification to compute the clusters [30]. In our experiments we will use the given clustering when possible, or the standard clustering techniques associated with the three existing frameworks (SOM, GMap, and TMFDL).

4.5 Overlap Removal

After the textual reduction, dimensionality reduction, and clustering steps the output is a collection of n points in 2D space, each of which is associated with a term from the input text. For many real-world data sets, the result of such pre-processing contains several dense clusters of points. Since we are interested in effectively communicating the underlying data, we need to not only show n points, but to print the text associated with them – the word-phrases associated with the top n terms from the input. Note that the existing three frameworks use different ad hoc solutions to this problem: both the SOM-based and the TMFDL-based approaches show only a subset of the labels that do not overlap, while the GMap-based approach limits the number of terms to be shown and uses post-processing to remove overlaps.

Before we can compare the effectiveness of point cloud, node-link, and map-like representations, we need to ensure that the text is readable. With this in mind we will design, implement, and evaluate several methods for consistent removal of the text overlaps and select the best one. We make two assumptions about this process: (1) all n terms will be printed and no pair of terms visually overlap, and (2) each term is assigned to a unique cluster.

Note that the goal of displaying all text in a non-overlapping fashion is a non-trivial task. For example, an overlap removal process that minimizes changes in the relative positions of the points is not necessarily the best, as it might break-up existing clusters, or severely affect the compactness of the layout. Conversely, by simply scaling the layout up we can remove all overlaps and not affect distortion; however, this is achieved at the expense in a very large blow-up in size, which affects compactness, as well as the ability to see the layout in its entirety. Here we describe two possible algorithms for overlap removal: *Inflate* and *Carve*.

In the Inflate algorithm we begin by placing all labels in the positions given by Phase 1. Then we scale down the entire drawing canvas until there are no overlaps. Note that this can always be done is no two input points have identical coordinates (if such points exist in the input, we perturb them slightly). The

resulting layout is free of overlaps but the labels a likely very tiny and the drawing canvas is huge. We improve the layout by iteratively increasing dimensions of all rectangles by some fixed up-scaling factor (hence "inflating" labels). After each iteration some labels may overlap. We resolve the overlaps using several iterations of a force-directed algorithm, where repulsive forces between labels push overlapping labels apart. Since the dimensions of each label grow by only a small factor (say 5%), these forces will likely preserve the relative positions of the labels.

The Carve Algorithm uses seam-carving, a content-aware image resizing technique, to keep the relative positions of labels unchanged. Here as in the previous algorithm, we begin with a preliminary overlap-free. Then the drawing canvas is divided into regions, and for each region an energy function is computed. A connected left-to-right or top-to-bottom path of low energy regions is called a seam. The major step of the algorithm is iteratively carving out seams of low energy to remove empty spaces between labels. Since the order of seam removals greatly affects the final result, a dynamic programming approach can be used to find the best such order.

Since we will use the same node placement in all three visualizations (point cloud, node-link, map) we only need to describe how to obtain the map-like representations. It is not easy to extract just the map-making facilities from the SOM framework; however, it is very easy to use the GMap map-making features. Once data is embedded and clustered (as in our case) GMap uses modified Voronoi diagrams to group nodes in the same clusters into countries and a coloring method which uses light pastel colors for the countries and ensures that neighbors have sufficiently different hues.

4.6 Quantitative Validation

Each of the proposed methods embeds a graph $G(V, E)$ with n vertices $V = \{v_i : 1 \leq i \leq n\}$ (where each vertex v_i corresponds to term t_i) onto a set of n points $P = \{p_i : 1 \leq i \leq n\}$ in the Euclidean plane. Each vertex v_i is assigned an xy-coordinate (x_i, y_i) to be embedded at point $p_i \in P$. To do this, each edge $e_{ij} = (v_i, v_j) \in E$ is assigned a weight $w_{ij} = \hat{S}(t_i, t_j)$, the rescaled similarity, of a pair terms These weights are then converted into ideal distances using a logarithmic transformation such as $d_{ij} = D(t_i, t_j) = -\log[(1 - s) \cdot \hat{S}(t_i, t_j) + s)]$, where s is a small positive constant.

One can evaluate how accurate such as embedding of a point set P is using one of three metrics: *stress*, *distortion*, or *precision and recall*. Stress, which is the evaluation function used for MDS, is defined as $stress(P) = \sum_{i<j}[(||p_i - p_j|| - d_{ij})^2/d_{ij}^2]$, where $||p_i - p_j||$ is the Euclidean distance between points p_i and p_j and d_{ij} is the ideal distance between each pair. Distortion as well as precision and recall are simpler measures that can be applied to unweighted graphs (where an edge is either in the graph or not). Distortion can be calculated using the same formula as stress, except that the graph theoretical shortest-path distance is used instead of the ideal distance d_{ij}. For a given k, the precision of point p_i in an embedding is defined as ratio of points within radius $r_i^k = ||p_i - p_i^k||$, where p_i^k is the k-nearest neighbor of point p_i, which correspond to neighbors to vertex v_i. Conversely, the recall of point p_i is the ratio of outside of radius r_i^k that are not neighbors of v_i. The precision and recall of point set P is then the average of the precision and recall of each point $p_i \in P$, respectively.

Other quantitative measures of embedding include the notions of compactness and uniform distribution. Compactness can be defined as the ratio of the total area of all labels (where each vertex has a corresponding label) over the bounding box of layout. When comparing two layouts where each of the labels of each vertex has the same size, then the ratio of their bounding boxes is another measure for compactness. The layout with the smaller bounding box is clearly more compact. One simple measure for uniform distribution is to compute the entropy of the points in the embedding, in how closely they match a uniform embedding as given by $\sum_{i \neq j} p_{ij} \log p_{ij}$ where p_{ij} is the number of points in the cell in the i^{th} row and the j^{th} column of a vnvn grid that overlays the bounding box of the embedding.

Additionally, each of the proposed methods clusters the terms into disciplines, for which modularity is one standard measure, namely, $Q = 1/2m \sum_{1 \leq i,j \leq n} [w_{ij} - 1/2m \cdot \sum_{1 \leq i \leq n} w_{ij} \cdot \sum_{1 \leq j \leq n} w_{ij}] \delta(C_i, C_j)$,

where $m = 1/2 \sum_{1 \leq i,j \leq n} w_{ij}$ is the sum of all edge weights, C_i and C_j are the assigned clusters (in this case disciplines) of v_i and v_j, respectively, and $\delta(C_i, C_j) = 1$ if and only if $i = j$. If one knows what the ideal clustering should be, then one can compare the compute clustering against the ideal by evaluating the coherence of each cluster, which is maximum percentage of vertices in a computed cluster that share the same ideal cluster.

4.7 Qualitative Validation

The output of dimensionality reduction are two-dimensional topic spaces that assign each document to a spatial position. Interested to understand the strengths and weaknesses of the different dimensionality reductions, novice users and experts will participate in studies to help us answer:

- Which reduction has the highest local and global accuracy? Does the spatial document clusters match how human experts organize/group documents?
- What document density and clustering are best? Assigning all documents in a cluster the same spatial position makes it impossible to see the number of documents per cluster. A uniform distribution of documents makes it hard to see clusters, see example in Fig. 7
- Are white spaces beneficial for recognition and memorization of cluster shapes and boundaries, (see example in Fig. 7)?

The work will not only build on prior studies [127], but will extend them to the first formal evaluation of widely used dimensionality reduction methods. A total of 120 subjects will be run. The general procedure is as follows: After giving consent, participants will (i) review complete a pre-test questionnaire that collects demographic info (age, gender, native language, expertise), (ii) read an information sheet with basic information on textual analysis and spatial layouts, and (iii) examine different Phase 1 results and answer associated questions. The participants will also be given the opportunity to provide additional feedback post-test questionnaires, and ask questions in a debriefing.

5 Planned Work—Phase 2: Topic Visualization

Phase 2 aims to compare different ways to visualize topic maps and data overlays, see general workflow in Fig. 6. Quantitative and qualitative measures will be used to validate and compare resulting visualizations.

Initial informal experiments indicate that the data from Figure **??** presented as a map attracts more viewers and holds the viewer's attention longer than the same data presented as a graph. It is possible that the familiarity with maps, or image is responsible for the greater interest and longer viewing times, but we believe that it is the "less boring" aspect of the map visualization compared to the node-and-link graph visualization that is at work here. We would like to test this formally, with the help of a carefully designed user study that would test whether maps indeed offer a better way for representing the underlying data.

We also noticed that people tend to perform spontaneous knowledge discovery tasks when looking at maps, even without being asked to do so. This leads us to believe that presenting the underlying data as a map, might have tangible and measurable advantages over presenting the data as a graph. Typical experiments testing the effectiveness of one graph drawing algorithm over another include performing timed tasks, such as finding the shortest path between two vertices and finding the average degree of a set of vertices. The type of knowledge discovery that seems conducive to map visualization is different. We will work on carefully designing experiments that would test how well maps do at exploratory visualization (what is in the data and where) and as well as on targeted tasks, such as identifying vertices that would disconnect the graph (e.g., "gateway" vertices that connect neighboring countries).

5.1 Datasets Used

The input to Phase 2 will be the best results from Phase 2. In addition, we will utilize the UCSD Map of Science network layout (see Section 2.4) validated external to this proposal. Its current rendering as a net-

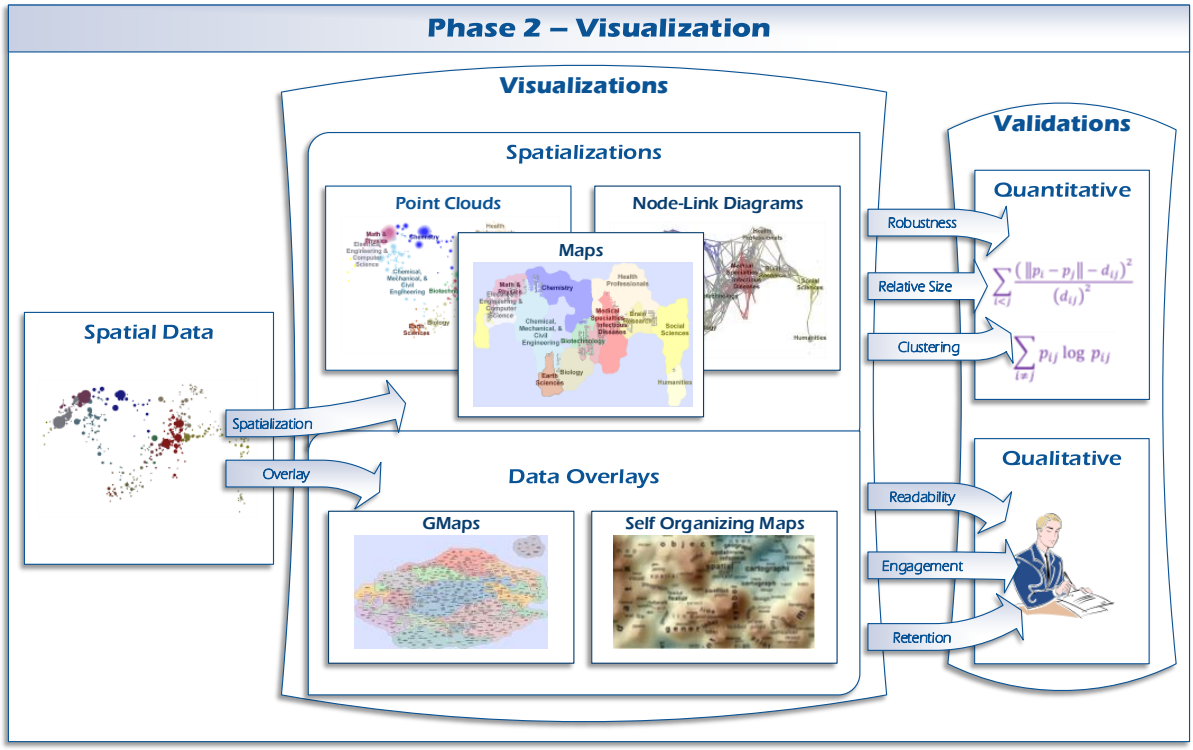

Figure 6: Phase 2—Visualization [[BOTTOM TO BE REPLACED WITH 4 DATA OVERLAYS]]

work layout will serve as a baseline, but it will also be re-rendered using existing techniques, to make it clear whether an improved result over baseline is the result of better data processing, or improved visualization techniques. In November 2013, a workshop will convene in Bloomington, IN that bring together researchers interested to create a new iteration of the UCSD Map of Science using 20 years of Web of Scopus paper-level data but also new data from the Chinese Academy of Sciences as well as the SciELO database, including 953 Brazilian journals, 600 of which have never been indexed in the map before.

5.2 Render Basemap and Data Overlays

The best workflows from Phase (and the UCSD Map of Science) will be rendered using different visual representations of the base map reference system: the point cloud, a node-link diagram of the same data, and a map-like representation, see Fig. 7.

One of the most common uses of a map is to overlay a dataset across it. Many types of data are overlaid across geographic maps including weather, traffic and election results. The ideal topical map would have similar utility, allowing a user to quickly encode their data and display it across the map in a way that others can read. Data overlays might be generated on the fly–in near real time. This project will generate and

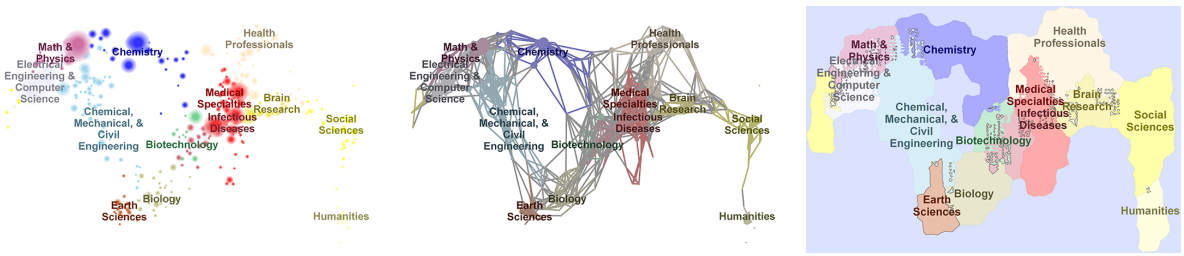

Figure 7: Point cloud, a node-link diagram, and map representation of the UCSD science map

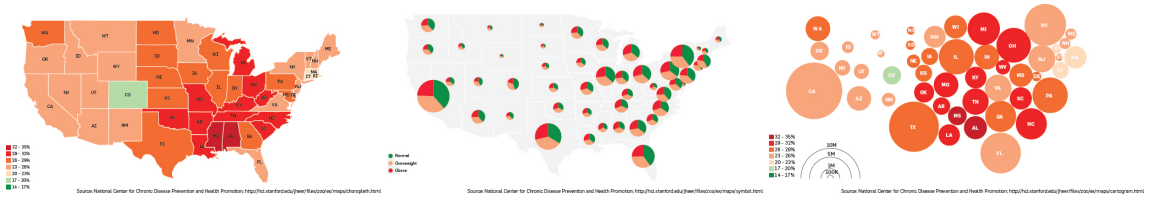

Figure 8: Data overlays for obesity rates in the US: choropleth, graduated symbols, Dorling cartogram (for the same data) from [85].

compare different data overlays, also called visual mappings, such as area size coding (cartogram), color coding (choropleth if using existing regions; heatmap if data defined regions), or proportional symbols that are easy to identify and compare, e.g., circles, squares, see Fig. 8.

Mappings of new records to existing base maps might be generated by mapping a new data record 1:1 or 1:n to a point or area. We hypothesize that 1:1 mappings of records to data points are easier to read while 1:n mappings to points or areas are harder to read. However, a 1:n mapping if additional data variables to multiple graphic symbol types, *e.g.*, symbol color and type might be easier to read than a single encoding.

5.3 Quantitative Validation

Once we have obtained a placement for the nodes that has no overlaps between their labels, *we will compare three visualizations (point cloud, node-link, map) that use exactly the same placement*. However, in the first one we only show the nodes and their labels, in the second one we also show the connections between the nodes, and in the last one we also explicitly color regions of the plane into countries. Here we would measure the potential of these visualizations to convey the information of the underlying data. Consider the following problems:

- Which area/cluster/region contains the most points? Here we anticipate that point clouds might be best.
- Which of the following two pairs of clusters are more tightly interconnected? Here we anticipate that the node-link and/or map representations would be better than the point cloud.
- Which of the following two clusters is bigger? Here it is possible that map representation might be best.

5.4 Qualitative Validation

In addition to the goals of a readable and memorable map, the study also seeks to create a map that is usable in a number of ways.

Human subject studies will be conducted to advance our understanding of what visualizations can be correctly "read" and used to support daily decision making by expert user and by a general audience. The work will benefit from user studies currently conducted in five science museums in the U.S. as part of the Pathways: Sense-Making of Big Data NSF project. Specifically, we plan to run user studies to identify the best (needs to be formally defined) base map, data overlay (including labeling), and mapping of new data on map for a given target audience and user need. Specifically, we will test visualizations for the following:

- *readability* (can users find certain areas, structures, pathways on map),
- *memorability* (can users answer if certain areas were covered and where they are on the map?),
- *reproducibility* (users are asked to study the map and then redraw it), and
- *usability* (are the maps and data overlays useful for human decision making).

6 Education Impact

The proposed research component is integrated with an educational component aiming to involve undergraduate, graduate, and postgraduate students through project-oriented courses. Building on prior teaching experience and published work in the area of computer science education [34, 35, 43–45], we focus on graduate student advising, and the involvement of students at all levels in research. International tutorials and research presentations will be given at major science of science, information visualization, and graph theory conferences. Innovative and effective approaches to science teaching will be formulated and disseminated using the Information Visualization MOOC (`http://ivmooc.cns.iu.edu`) that has registered students from more than 100 countries. As in prior projects, high-quality, validated datasets, as well as high-quality, functional software developed for this project will be made available to the broader research community in support of teaching and replication of results and future algorithm comparisons. All our new contributions will become a part of the Sci2 plug-and-play macroscope collection [18].

The PIs have already organized several Dagstuhl Seminars, including Dagstuhl Seminar 12261 "Putting Data on the Map" and Dagstuhl Seminar 13151 "Drawing Graphs and Maps with Curves". These week-long, by-invitation-only meetings are subsidized by the German Science Foundation and aim to promote computer science research by the transfer of knowledge between the research and application side of informatics, tapping into new fields of application, and fostering the next generation of researchers by including them in the research dialogue. The PIs will co-organize a multi-disciplinary Dagstuhl Seminar on the topic of "Mapping Science" with participants from information visualization, network visualization, big data analysis, geography, and GIS. In addition to experts in the field, we will have participants from industry (e.g., publishers such as Elsevier and Springer, data visualization companies such as IBM and Tableau). Undergraduate, graduate, and postgraduate students are essential participants in these events. Previous participants have had their careers changed by finding research topics that became their PhD thesis, or new colleagues who became postdoctoral advisors.

7 Prior NSF Support

Börner's team was/is involved in a number of NSF-funded projects. Major research contributions and publications are listed here.

SEAD Sustainable Environment Through Actionable Data. NSF OCI-0940824 DataNet award (Margaret Hedstrom, Myron P. Gutmann, Praveen Kumar, Jim Myers, and Beth Plale). Robert P. Light implemented a VIVO (http://vivoweb.org) driven marketplace that sustainability researchers can use to share research data.

Pathways: Sense-Making of Big Data. NSF ISE DRL-1223698 Award (Katy Börner, Adam V. Maltese, Joe E. Heimlich, Stephen Miles Uzzo, Paul Martin, and Sasha Palmquist). This recently funded project aims to identify what data visualizations a general audience can read. Two studies have been conducted interviewing and testing more than 600 adults and youth at six U.S. science museums. Publications are forthcoming.

Digging by Debating: Linking Massive Datasets to Specific Arguments. Digging Into Data - NSF, NEH, JISC Award (Colin Allen, Katy Börner, Chris Reed, Andrew Ravenscroft, and David Bourget). This US-UK collaboration develops a multi-scale workbench, called "InterDebates", that helps extract and visualize argumentative structures from large datasets using a mixture of automated and social computing techniques. Publication is forthcoming.

SGER/Collaborative Research: Mapping the Structure and Evolution of Sustainability Science Research. NSF CBET-0831636 (Katy Börner, and Luis M. A. Bettencourt). This project implemented an online interactive interface to publication, patent, and grant datasets, see `http://mapsustain.cns.iu.edu` and [23, 28, 130].

SEI: NetWorkBench: A Large-Scale Network Analysis, Modeling and Visualization Toolkit for Biomedical, Social Science and Physics Research. NSF IIS-0513650 Award (Katy Börner, Albert-Lszl

Barabsi, Santiago Schnell, Alessandro Vespignani, Stanley Wasserman, and Eric Wernert). This project resulted in the design of the first plug-and-play macroscope: the Network Workbench, available at http://nwb.cns.iu.edu, and its application in science of science research, see [14–16, 18, 20, 21, 25, 26, 49, 87–90, 94, 109, 114, 118–120, 133].

Creative Metaphors to Stimulate New Approaches to Visualizing, Understanding, and Rethinking Large Repositories of Scholarly Data. NSF IIS-0715303 Award (Katy Börner). Among others, this project resulted in a comic book (see `http://scimaps.org/exhibit/docs/ComicBook_web.pdf`) and the Humanexus movie (see `http://yfshen.info/humanexus/`) that won three awards and enjoys many official festival selections and screenings, see [21, 24, 28, 86, 133, 139].

Collaborative Research: Social Networking Tools to Enable Collaboration in the Tobacco Surveillance, Epidemiology, and Evaluation Network (TSEEN). Collaborative Systems NSF IIS-0534909 Grant (Katy Börner, Thomas Finholt, and Gary Giovino), see [18, 20, 21, 86, 120]. **TLS: Towards a Macroscope for Science Policy Decision Making.** NSF SBE-0738111 Award (Katy Börner, Weixia (Bonnie) Huang, Kevin W. Boyack, Mark Price, Lokman Meho, Micah Linnemeier, Qizheng (Stanley) Bao, and Shreyas Ahir , $399,870) 2008.01.01 - 2009.12.31, see [8, 9, 18, 23, 24, 27, 28, 86, 109, 117, 138].

Kobourov's team was/is involved in a number of NSF-funded projects. Major research contributions and publications are listed here.

Visualization of Giga-Graphs and Graph Processes, (PI: Kobourov), NSF ACR, 2002-05, $240,358: This project has resulted in a number of publications on visualization of large graphs [32, 42, 54, 57, 59–61, 72] and several software systems for graph visualization, including GraphAEL [58] (for visualization of computing literature) and GMorph [62] (for intersection-free morphing of planar graphs). Many students have been involved in this research, both of the graduate and undergraduate levels. Three PhD students completed PhD theses on work funded by this grant. Cesim Erten completed a PhD thesis [56] with work funded by this grant [32, 55, 57–63, 95]. More than a dozen of our publications have at least one undergraduate co-author [37, 39–43, 54, 57, 58, 61, 72, 100]. Four MS students were co-authors on papers related to the project [1, 37, 42, 57, 58, 62, 63, 72, 101, 103]. Undergraduates who have worked on this project have won several awards Gary Yee (Outstanding Senior, UA Department of Computer Science, 2004. Outstanding Senior, UA College of Science, 2004. CRA Outstanding Undergraduate Award, 2004), Kyriakos Pavlou (Outstanding Senior, UA Dept. of Computer Science, 2005. Outstanding Senior, UA College of Science, Fall 2005), Kevin Wampler (Galileo Circle Award, UA College of Science, 2005)

CAREER: Embedding, Morphing, and Visualizing Dynamic Graphs, (PI: Kobourov), NSF-CCF, 2006-11, $419,645: This project has resulted in a number of publications on visualization of large graphs [7, 33, 78] and graphs that evolve through time [93, 115], several software systems for graph visualization, including MoCS [80] (Maps of Computer Science), GMap [83, 91], GraphSET [68], and Lombardi [38]. Book chapters on force-directed algorithms [99], simultaneous embedding [12], and map-based visualization [84] also resulted from this work. New graph visualization models such as Lombardi graph drawing were proposed [50, 51, 124]. Joe Fowler co-authored 8 papers [31, 65–67, 74–77] and completed a PhD in 2009 [73]. Alejandro Estrella-Balderrama co-authored 8 papers [10, 36, 65–70] and completed a PhD in 2009 [64]. Jan-Hinrich Kämper (MS) co-authored a paper [98]. Undergraduates who have worked on this project have won several awards: Anand Iyer (Outstanding Senior, UA Dept. of Computer Science, 2006), David Forester (Outstanding Senior, UA Dept. of Computer Science, 2007), Daniel Fried (Barry Goldwater Scholar), Katherine Cunningham (Outstanding Senior, UA College of Science 2012).

Algorithms for Visualizing Data with Contact Graphs, (PI: Kobourov), NSF-CCF, 2010-13, $296,001: This project focuses on the theoretical aspects of map representations. Several papers in computational geometry and graph theory [52, 53, 71, 102] resulted from this work. One Two PhD students, one MS student, and three undergraduate research assistants were partly supported by this grant. Jawaherul Alam's PhD thesis is based on a series of papers done as a part of this project [4–6]. Sankar Veeramoni's PhD thesis is also based on work done as a part of this research project [92, 93]. Jackson Toeniskoetter's MS thesis is on the

topic of threshold graph labeling.

Collaborative: ImageQuest: Calibrated Imaging and Validated Analysis, (PIs: Cindy Grimm, Stephen Kobourov, Jarlath O'Neil, Robert Pless, Ruth West), NSF-IBIV, 2011-14, $1,268,593: This grant is a collaboration between researchers at Washington University, University of Vermont, University of Arizona and University of California-San Diego. The major goals of the project are to research mechanisms through which citizen science contributions to biological imaging projects are useful, by providing tools to motivate ongoing, continued engagement in projects and to make possible more quantitative measurements through calibrated data capture and data annotation. In total, this project is supporting 5 PhD students, 2 postdoctoral students, and 10 undergraduate research assistants. More than a dozen publications [2, 3, 46, 106, 110, 113, 134–137] have been published and several more are under submission [47, 123, 132].

References

[1] J. Abello, S. G. Kobourov, and R. Yusufov. Visualizing large graphs with compound-fisheye views and treemaps. In *12th Symposium on Graph Drawing (GD)*, pages 431–441, 2005.

[2] A. Abrams, C. Hawley, and R. Pless. Heliometric stereo: shape from sun position. In *Computer Vision–ECCV 2012*, pages 357–370. 2012.

[3] A. Abrams and R. Pless. Web-accessible geographic integration and calibration of webcams. *ACM Trans. Multimedia Comput. Commun. Appl.*, 9(1):8:1–8:20, Feb. 2013.

[4] M. J. Alam, T. C. Biedl, S. Felsner, A. Gerasch, M. Kaufmann, and S. G. Kobourov. Linear-time algorithms for hole-free rectilinear proportional contact graph representations. *Algorithmica*, 67(1):3–22, 2013.

[5] M. J. Alam, T. C. Biedl, S. Felsner, M. Kaufmann, and S. G. Kobourov. Proportional contact representations of planar graphs. *J. Graph Algorithms Appl.*, 16(3):701–728, 2012.

[6] M. J. Alam, T. C. Biedl, S. Felsner, M. Kaufmann, S. G. Kobourov, and T. Ueckerdt. Computing cartograms with optimal complexity. *Discrete & Computational Geometry*, 50(3):784–810, 2013.

[7] M. A. Bekos, M. Kaufmann, S. G. Kobourov, and A. Symvonis. Smooth orthogonal layouts. In *20th Symposium on Graph Drawing (GD)*, pages 150–161, 2012.

[8] P. v. d. Besselaar, K. Börner, and A. Scharnhorst, editors. *Models of Science Dynamics: Encounters Between Complexity Theory and Information Science*. Springer Verlag, 2012.

[9] J. R. Biberstine, K. Börner, R. J. Duhon, E. F. Allgood, and A. Skupin. A semantic map of the last.fm music folksonomy. In *Seventh International Conference on Geographic Information Science (GIScience 2012)*, 2012.

[10] C. Binucci, E. D. Diacomo, W. Didimo, A. Estrella-Balderrama, F. Frati, S. G. Kobourov, and G. Liotta. Directed graphs with an upward straight-line embedding into every point set. In *21th Canadian Conference on Computational Geometry (CCCG)*, 2009.

[11] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python*. O'Reilly Media, Incorporated, 2009.

[12] T. Bläsius, S. G. Kobourov, and I. Rutter. Simultaneous embedding of planar graphs. In R. Tamassia, editor, *Handbook of Graph Drawing and Visualization*, pages 349–381. CRC Press, 2013.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[14] K. Börner. Mapping science. In *International Conference and Workshop on Network Science*, 2006.

[15] K. Börner. Science of science policy position paper. In *Workshop on Science of Science Policy: Developing our Understanding of Public Investments in Science*, 2006.

[16] K. Börner. Making Sense of Mankind's Scholarly Knowledge and Expertise: Collecting, Interlinking, and Organizing What we Know and Different Approaches to Mapping (network) science. *Environment and Planning B: Planning and Design*, 34:808–825, 2007.

[17] K. Börner. *Atlas of Science: Visualizing What We Know*. MIT Press, 2010.

[18] K. Börner. Plug-and-play macroscopes. *Communications of the ACM*, 54(3):60–69, Mar. 2011.

[19] K. Börner, C. Chen, and K. W. Boyack. Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37(1):179–255, 2003.

[20] K. Börner, E. Hardy, B. Herr, T. Holloway, and W. B. Paley. Taxonomy visualization in support of the semi-automatic validation and optimization of organizational schemas. *Journal of Informetrics*, 1(3):214 – 225, 2007.

[21] K. Börner, W. Huang, M. Linnemeier, R. Duhon, P. Phillips, N. Ma, A. Zoss, H. Guo, and M. Price. Rete-netzwerk-red: analyzing and visualizing scholarly networks using the network workbench tool. *Scientometrics*, 83(3):863–876, 2010.

[22] K. Börner, R. Klavans, M. Patek, A. M. Zoss, J. R. Biberstine, R. P. Light, V. Larivière, and K. W. Boyack. Design and update of a classification system: The ucsd map of science. *PLoS ONE*, 7(7):e39464, 07 2012.

[23] K. Börner, N. Ma, R. J. Duhon, and A. Zoss. Science and technology assessment using open data and open code. *IEEE Intelligent Systems*, 24(4):78–81, 2009.

[24] K. Börner, F. Palmer, J. M. Davis, E. F. Hardy, S. M. Uzzo, and B. J. Hook. Teaching children the structure of science. In *SPIE Conference on Visualization and Data Analysis*, volume 7243, pages 1–14, 2009.

[25] K. Börner, S. Penumarthy, M. Meiss, and W. Ke. Mapping the diffusion of scholarly knowledge among major u.s. research institutions. *Scientometrics*, 68(3):415–426, 2006.

[26] K. Börner, S. Sanyal, and A. Vespignani. Network science. *Annual Review of Information Science and Technology*, 41(1):537–607, 2007.

[27] K. Börner and A. Scharnhorst. Visual conceptualizations and models of science. *Journal of Informetrics*, 3(3):161 – 172, 2009. ¡ce:title¿Science of Science: Conceptualizations and Models of Science¡/ce:title¿.

[28] K. Börner, R. M. Wagner, N. Ma, J. R. Biberstine, R. Berhane, H. Jiang, S. E. Ivey, K. Pearson, , and C. McCabek. Introducing the science of science (sci2) tool to the reporting branch, office of extramural research/office of the director, national institutes of health, 2010. Workshop on the Science of Science Measurement, December 2-3, Washington D.C.

[29] K. Boyack, K. Börner, and R. Klavans. Mapping the structure and evolution of chemistry research. *Scientometrics*, 79(1):45–60, 2007.

[30] K. W. Boyack, R. Klavans, and K. Börner. Mapping the backbone of science. *Scientometrics*, 64(3):351–374, 2005.

[31] U. Brandes, C. Erten, J. J. Fowler, F. Frati, M. Geyer, C. Gutwenger, S.-H. Hong, M. Kaufmann, S. G. Kobourov, G. Liotta, P. Mutzel, and A. Symvonis. Colored simultaneous geometric embeddings. In *13th Conference on Computing and Combinatorics (COCOON)*, pages 254–263, 2007.

[32] P. Brass, E. Cenek, C. A. Duncan, A. Efrat, C. Erten, D. Ismailescu, S. G. Kobourov, A. Lubiw, and J. S. B. Mitchell. On simultaneous graph embedding. *Computational Geometry: Theory and Applications*, 36(2):117–130, 2007.

[33] D. Bremner, W. S. Evans, F. Frati, L. J. Heyer, S. G. Kobourov, W. J. Lenhart, G. Liotta, D. Rappaport, and S. Whitesides. On representing graphs by touching cuboids. In *20th Symposium on Graph Drawing (GD)*, pages 187–198, 2012.

[34] S. Bridgeman, M. T. Goodrich, S. G. Kobourov, and R. Tamassia. `PILOT`: An interactive tool for learning and grading. In *Proceedings of the 31st Technical Symposium on Computer Science Education (SIGCSE 2000)*, pages 139–143, 2000.

[35] S. Bridgeman, M. T. Goodrich, S. G. Kobourov, and R. Tamassia. `SAIL`: A system for generating, archiving, and retrieving specialized assignments using LaTex. In *Proceedings of the 31st Technical Symposium on Computer Science Education (SIGCSE 2000)*, pages 300–304, 2000.

[36] J. Cappos, A. Estrella-Balderrama, J. J. Fowler, and S. G. Kobourov. Simultaneous graph embedding with bends and circular arcs. *Comput. Geom.*, 42(2):173–182, 2009.

[37] J. Cappos, S. G. Kobourov, M. Miles, M. Stepp, K. Pavlou, and A. Wixted. Collaboration with diamondtouch. In *10th International Conference on Human Computer Interaction (INTERACT)*, pages 986–989, 2005.

[38] R. Chernobelskiy, K. I. Cunningham, M. T. Goodrich, S. G. Kobourov, and L. Trott. Force-directed Lombardi-style graph drawing. In *Graph Drawing*, pages 320–331, 2011.

[39] C. Collberg, S. G. Kobourov, E. Carter, and C. Thomborson. Error-correcting graphs for software watermarking. In *29th Workshop on Graph Theoretic Concepts in Computer Science*, pages 156–167, 2003.

[40] C. Collberg, S. G. Kobourov, S. Kobes, B. Smith, S. Trush, and G. Yee. Tetratetris: An application of multi-user touch-based human-computer interaction. In *9th International Conference on Human-Computer Interaction (INTERACT)*, pages 81–88, 2003.

[41] C. Collberg, S. G. Kobourov, J. Louie, and T. Slattery. SPLAT: A system for self-plagiarism detection. In *Proceedings of the IADIS Conference WWW/Internet*, pages 508–514, 2003.

[42] C. Collberg, S. G. Kobourov, J. Nagra, J. Pitts, and K. Wampler. A system for graph-based visualization of the evolution of software. In *ACM Symposium on Software Visualization (SoftVis)*, pages 77–86, 2003.

[43] C. Collberg, S. G. Kobourov, and S. Westbrook. Algovista: A tool to enhance algorithm design and understanding. In *7th Symposium on Innovation and Technology in Computer Science Education (ITICSE)*, pages 228–237, 2002.

[44] C. S. Collberg, S. Debray, S. G. Kobourov, and S. Westbrook. Increasing undergraduate involvement in computer science research. In *8th World Conference on Computers in Education (WCCE)*, pages 342–352, 2005.

[45] C. S. Collberg, S. G. Kobourov, and S. Westbrook. Algovista: an algorithmic search tool in an educational setting. In *35th Symposium on Computer Science Education (SIGCSE)*, pages 462–466, 2004.

[46] A. Das, E. R. Gansner, M. Kaufmann, S. G. Kobourov, J. Spoerhase, and A. Wolff. Approximating minimum manhattan networks in higher dimensions. In *19th European Symposium on Algorithms (ESA)*, pages 49–60, 2011.

[47] L. De La Cruz, S. Kobourov, S. Pupyrev, P. Shen, and S. Veeramoni. Angryants: An approach for accurate average trajectories using citizen science. *arXiv preprint arXiv:1212.0935*, 2012.

[48] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[49] R. Duhon. Understanding outside collaborations of the chinese academy of sciences using jensen-shannon divergence. In *Proc. SPIE, Visualization and Data Analysis*, volume 7243, pages 72430C–72430C–7, 2009.

[50] C. A. Duncan, D. Eppstein, M. T. Goodrich, S. G. Kobourov, and M. Löffler. Planar and poly-arc Lombardi drawings. In *Graph Drawing*, pages 308–319, 2011.

[51] C. A. Duncan, D. Eppstein, M. T. Goodrich, S. G. Kobourov, and M. Nöllenburg. Lombardi drawings of graphs. *J. Graph Algorithms Appl.*, 16(1):85–108, 2012.

[52] C. A. Duncan, D. Eppstein, M. T. Goodrich, S. G. Kobourov, and M. Nöllenburg. Drawing trees with perfect angular resolution and polynomial area. *Discrete & Computational Geometry*, 49(2):157–182, 2013.

[53] C. A. Duncan, E. R. Gansner, Y. F. Hu, M. Kaufmann, and S. G. Kobourov. Optimal polygonal representation of planar graphs. *Algorithmica*, 63(3):672–691, 2012.

[54] B. Dux, A. Iyer, S. Debray, D. Forrester, and S. G. Kobourov. Visualizing the behaviour of dynamically modifiable code. In *13th IEEE Workshop on Porgram Comprehension*, pages 337–340, 2005.

[55] A. Efrat, C. Erten, and S. G. Kobourov. Fixed-location circular-arc drawing of planar graphs. In *11th Symposium on Graph Drawing*, pages 147–158, 2003.

[56] C. Erten. *Simultaneous Embedding and Visualization of Graphs*. PhD thesis, University of Arizona, 2004.

[57] C. Erten, P. J. Harding, S. Kobourov, K. Wampler, and G. Yee. Exploring the computing literature using temporal graph visualization. In *Visualization and Data Analysis*, pages 45–56, 2004.

[58] C. Erten, P. J. Harding, S. G. Kobourov, K. Wampler, and G. Yee. GraphAEL: Graph animations with evolving layouts. In *11th Symposium on Graph Drawing*, pages 98–110, 2003.

[59] C. Erten and S. G. Kobourov. Simultaneous embedding of a planar graph and its dual on the grid. *Theory of Computing Systems*, 38(3):313–327, 2005.

[60] C. Erten and S. G. Kobourov. Simultaneous embedding of planar graphs with few bends. *Journal of Graph Algorithms and Applications*, 9(3):347–364, 2005.

[61] C. Erten, S. G. Kobourov, A. Navabi, and V. Le. Simultaneous graph drawing: Layout algorithms and visualization schemes. *Journal of Graph Algorithms and Applications*, 9(1):165–182, 2005.

[62] C. Erten, S. G. Kobourov, and C. Pitta. Intersection-free morphing of planar graphs. In *11th Symposium on Graph Drawing*, pages 320–331, 2003.

[63] C. Erten, S. G. Kobourov, and C. Pitta. Morphing planar graphs. In *20th ACM Symposium on Computational Geometry*, 2004. To appear in 2004.

[64] A. Estrella-Balderrama. *Simultaneous Embedding and Level Planarity*. PhD thesis, University of Arizona, 2009.

[65] A. Estrella-Balderrama, J. J. Fowler, and S. G. Kobourov. Characterization of unlabeled level planar trees. In *14th Symposium on Graph Drawing (GD)*, pages 367–379, 2006.

[66] A. Estrella-Balderrama, J. J. Fowler, and S. G. Kobourov. Colored simultaneous geometric embeddings and universal pointsets. In *21th Canadian Conference on Computational Geometry (CCCG)*, 2009.

[67] A. Estrella-Balderrama, J. J. Fowler, and S. G. Kobourov. On the characterization of level planar trees by minimal patterns. In *17th Symposium on Graph Drawing (GD)*, 2009.

[68] A. Estrella-Balderrama, J. J. Fowler, and S. G. Kobourov. Graphset, a tool for simultaneous graph drawing. *Software Practice and Experience*, 40(10):849–863, 2010.

[69] A. Estrella-Balderrama, F. Frati, and S. G. Kobourov. Upward straight-line embeddings of directed graphs into point sets. In *34th Workshop on Graph-Theoretic Concepts in Computer Science (WG)*, pages 122–133, 2008.

[70] A. Estrella-Balderrama, E. Gassner, M. Jünger, M. Percan, M. Schaefer, and M. Schulz. Simultaneous geometric graph embeddings. In *15th Symposium on Graph Drawing (GD)*, pages 280–290, 2007.

[71] W. S. Evans, S. Felsner, M. Kaufmann, S. G. Kobourov, D. Mondal, R. I. Nishat, and K. Verbeek. Table cartograms. In *21st European Symposium on Algorithms (ESA)*, pages 421–432, 2013.

[72] D. Forrester, S. G. Kobourov, A. Navabi, K. Wampler, and G. Yee. graphael: A system for generalized force-directed layouts. In *12th Symposium on Graph Drawing (GD)*, 2004.

[73] J. J. Fowler. *Unlabeled Level Planarity*. PhD thesis, University of Arizona, 2009.

[74] J. J. Fowler, C. Gutwenger, M. Jünger, P. Mutzel, and M. Schulz. An spqr-tree approach to decide special cases of simultaneous embedding with fixed edges. In *16th Symposium on Graph Drawing (GD)*, pages 157–168, 2008.

[75] J. J. Fowler, M. Jünger, S. G. Kobourov, and M. Schulz. Characterizations of restricted pairs of planar graphs allowing simultaneous embedding with fixed edges. *Comput. Geom.*, 44(8):385–398, 2011.

[76] J. J. Fowler and S. G. Kobourov. Characterization of unlabeled level planar graphs. In *15th Symposium on Graph Drawing (GD)*, pages 37–49, 2007.

[77] J. J. Fowler and S. G. Kobourov. Minimum level nonplanar patterns for trees. In *15th Symposium on Graph Drawing (GD)*, pages 69–75, 2007.

[78] J. J. Fowler and S. G. Kobourov. Planar preprocessing for spring embedders. In *20th Symposium on Graph Drawing (GD)*, pages 388–399, 2012.

[79] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130, 2000.

[80] D. Fried and S. G. Kobourov. Maps of computer science. *arXiv preprint arXiv:1304.2681*, 2013.

[81] P. Gajer, M. T. Goodrich, and S. G. Kobourov. A fast multi-dimensional algorithm for drawing large graphs. *Computational Geometry: Theory and Applications*, 29(1):3–18, 2004.

[82] P. Gajer and S. G. Kobourov. GRIP: Graph dRawing with Intelligent Placement. *Journal of Graph Algorithms and Applications*, 6(3):203–224, 2002.

[83] E. Gansner, Y. Hu, S. Kobourov, and C. Volinsky. Putting recommendations on the map - visualizing clusters and relations. In *Proc. 3rd ACM Conference on Recommender Systems*, pages 345–354, 2009.

[84] E. Gansner, Y. Hu, and S. G. Kobourov. Viewing abstract data as maps. In T. Huang, editor, *Human Centric Visualization: Theories, Methodologies and Case Studies*, pages 63–93. Springer, 2013.

[85] J. Heer, M. Bostock, and V. Ogievetsky. A tour through the visualization zoo. *Commun. ACM*, 53(6):59–67, 2010.

[86] B. Herr, R. Duhon, K. Börner, E. Hardy, and S. Penumarthy. 113 years of physical review: Using flow maps to show temporal and topical citation patterns. In *Information Visualisation, 2008. IV '08. 12th International Conference*, pages 421–426, 2008.

[87] B. Herr, W. Ke, E. Hardy, and K. Börner. Movies and actors: Mapping the internet movie database. In *Information Visualization, 2007. IV '07. 11th International Conference*, pages 465–469, 2007.

[88] B. Herr, E. Talley, G. Burns, D. Newman, and G. LaRowe. The NIH visual browser: An interactive visualization of biomedical research. In *Information Visualisation, 2009 13th International Conference*, pages 505–509, 2009.

[89] B. W. Herr, W. Huang, S. Penumarthy, and K. Börner. Designing highly flexible and usable cyberinfrastructures for convergence. *Annals of the New York Academy of Sciences*, 1093(1):161–179, 2006.

[90] T. Holloway, M. Boičević, and K. Börner. Analyzing and visualizing the semantic coverage of wikipedia and its authors: Research articles. *Complexity, Special Issue on Understanding Complex Systems*, 12(3):30–40, Jan. 2007.

[91] Y. Hu, E. Gansner, and S. Kobourov. Visualizing Graphs and Clusters as Maps. *IEEE Computer Graphics and Applications*, 30(6):54–66, 2010.

[92] Y. Hu, S. G. Kobourov, and S. Veeramoni. On maximum differential graph coloring. In *18th Symposium on Graph Drawing (GD)*, pages 274–286, 2010.

[93] Y. Hu, S. G. Kobourov, and S. Veeramoni. Embedding, clustering and coloring for dynamic maps. In *5th IEEE PacificVis Symposium*, pages 33–40, 2012.

[94] W. Huang, B. W. H. II, S. Penumarthy, B. Markines, and K. Börner. Cishell - a plug-in based software architecture and its usage to design an easy to use, easy to extend cyberinfrastructure for network scientists. In *International Conference and Workshop on Network Science*, 2006.

[95] A. Iyer, A. Efrat, C. Erten, D. Forrester, and S. G. Kobourov. A force-directed approach to sensor localization. In *13th Symposium on Graph Drawing (GD)*, 2005.

[96] P. Jaccard. *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz, 1901.

[97] I. T. Jolliffe. *Principal Component Analysis*. Springer, second edition, October 2002.

[98] J. Kämper, S. G. Kobourov, and M. Nöllenburg. Circular-arc cartograms. In *6th IEEE PacificVis Symposium*, pages 1–9, 2013.

[99] S. G. Kobourov. Force-directed drawing algorithms. In R. Tamassia, editor, *Handbook of Graph Drawing and Visualization*, pages 383–408. CRC Press, 2013.

[100] S. G. Kobourov, A. Efrat, D. Forrester, and A. Iyer. Force-directed approaches to sensor network localization. In *8th Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 108–118, 2006.

[101] S. G. Kobourov and C. Pitta. An interactive multi-user system for simultaneous graph drawing. In *12th Symposium on Graph Drawing (GD)*, 2004.

[102] S. G. Kobourov, T. Ueckerdt, and K. Verbeek. Combinatorial and geometric properties of planar laman graphs. In *24th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1668–1678, 2013.

[103] S. G. Kobourov and K. Wampler. Non-Euclidean spring embedders. *IEEE Transactions on Visualization and Computer Graphics*, 11(6):757–767, 2005.

[104] T. Kohonen. *Self-organizing maps*. Springer, 2001.

[105] J. B. Kruskal and M. Wish. *Multidimensional Scaling*. Sage Press, 1978.

[106] S. Kurtek, J. Su, C. Grimm, M. Vaughan, R. Sowell, and A. Srivastava. Statistical analysis of manual segmentations of structures in medical images. *Computer Vision and Image Understanding*, 2013.

[107] G. LaRowe, S. Ambre, J. Burgoon, W. Ke, and K. Börner. The scholarly database and its utility for scientometrics research. *Scientometrics*, 79(2):219–234, 2009.

[108] M. Ley. DBLP - some lessons learned. *PVLDB*, 2(2):1493–1500, 2009.

[109] R. Light, T. Polley, and K. Börner. Open data and open code for big science of science studies. In *Proceedings of International Society of Scientometrics and Informetrics Conference*, pages 1342–1356, 2013.

[110] J. Little, A. Abrams, and R. Pless. Tools for richer crowd source image annotations. In *IEEE Workshop on Applications of Computer Vision (WACV),*, pages 369–374, 2012.

[111] S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian. Interactive, topic-based visual text summarization and analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 543–552, 2009.

[112] S. Lloyd. Last square quantization in PCM. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[113] S. MacFaden, J. O'Neil-Dunne, A. Royar, J. Lu, and A. Rundle. High-resolution tree canopy mapping for new york city using lidar and object-based image analysis. *Journal of Applied Remote Sensing*, 2013. Accepted, to appear in 2013.

[114] B. Markines. *Socially Induced Semantic Networks and Applications*. PhD thesis, Indiana University, Bloomington, 2009. Doctor of Philosophy in Computer Science: Filippo Menczer (Chair), Katy Börner, Randall Bramley and Dennis Groth.

[115] D. Mashima, S. G. Kobourov, and Y. Hu. Visualizing dynamic data with maps. *IEEE Trans. Vis. Comput. Graph.*, 18(9):1424–1437, 2012.

[116] A. K. McCallum. *MALLET: A Machine Learning for Language Toolkit*. 2002.

[117] S. Milojević, K. Börner, S. Morris, and K. W. Boyack. An introduction to modeling science: Basic model types, key definitions, and a general framework for the comparison of process models. In A. Scharnhorst, K. Börner, and P. Besselaar, editors, *Models of Science Dynamics*, Understanding Complex Systems, pages 3–22. Springer Berlin Heidelberg, 2012.

[118] C. Murray, W. Ke, and K. Börner. Mapping scientific disciplines and author expertise based on personal bibliography files. In *Information Visualization, 2006. IV 2006. Tenth International Conference on*, pages 258–263, 2006.

[119] C. Murray, W. Ke, H. Milanaov, M. Meiss, S. Rajagopal, and K. Börner. Geographical visualization of technology data in the US, 2005. IEEE InfoVis 2005 Contest Entry.

[120] T. Neirynck and K. Börner. Representing, analyzing, and visualizing scholarly data in support of research management. In *Information Visualization, 2007. IV '07. 11th International Conference*, pages 124–129, 2007.

[121] M. E. J. Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103:8577–8582, 2006.

[122] A. Noack. Moduarity clustering is force-directed layout. *Physical Review E*, 79, 2009.

[123] E. Packer, S. Pupyrev, A. Efrat, and S. Kobourov. Efficient methods for registration of multiple moving points in noisy environments. Technical Report TR13-01, Department of Computer Science, University of Arizona, 2013.

[124] H. C. Purchase, J. Hamer, M. Nöllenburg, and S. G. Kobourov. On the usability of Lombardi graph drawings. In *Graph Drawing*, pages 451–462, 2012.

[125] A. Skupin. A cartographic approach to visualizing conference abstracts. *IEEE Computer Graphics and Applications*, 22(1):50–58, 2002.

[126] A. Skupin. The world of geography: Visualizing a knowledge domain with cartographic means. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5274–5278, 2004.

[127] A. Skupin, J. R. Biberstine, and K. Börner. Visualizing the topical structure of the medical sciences: A self-organizing map approach. *PloS one*, 8(3):e58779, 2013.

[128] A. Skupin and C. de Jongh. Visualizing the ICA: A content-based approach. In *Proceedings of 22nd International Cartographic Conference*, 2005.

[129] P. Srinivasan. MeSHmap: a text mining tool for MEDLINE. *Proceedings AMIA Symposium*, pages 642–646, 2001.

[130] M. Stamper, C. H. Kong, N. Ma, A. Zoss, and K. Börner. chapter MAPSustain: Visualising biomass and biofuel research, pages 57–61. 2012.

[131] J. R. R. Tolkien. *The Shaping of Middle-Earth*. Houghton Mifflin Harcourt, 1986.

[132] M. Vaughan, C. Grimm, R. West, R. Sowell, R. Pless, and S. Kobourov. Specializing interfaces for citizen science segmentation of volumetric data. Technical Report 2012-42, Washington University St. Louis, Department of Computer Science and Engineering, 2012.

[133] E. A. Wernert, J. Lakshmipathy, M. Boyles, and K. Börner. Id2 – a scalable and flexible mixed-media information visualization system for public learning exhibits. In G. Siemens and C. Fulford, editors, *World Conference on Educational Multimedia, Hypermedia and Telecommunications 2009*, pages 3848–3856, Honolulu, HI, USA, 06/2009 2009. AACE.

[134] R. West, A. Halley, D. Gordon, J. O'Neil-Dunne, and R. Pless. Collaborative rephotography. In *ACM SIGGRAPH 2013 Studio Talks*, page 20. ACM, 2013.

[135] R. West, A. Halley, J. ONeil-Dunne, D. Gordon, and R. Pless. Collaborative imaging of urban forest dynamics: augmenting re-photography to visualize changes over time. In *IS&T/SPIE Electronic Imaging*, pages 86490L–86490L, 2013.

[136] R. West, T. Margolis, J. O'Neil-Dunne, and E. Mendelowitz. Metatree: augmented reality narrative explorations of urban forests. In *IS&T/SPIE Electronic Imaging*, 2012.

[137] R. West, T. Margolis, J. O'Neil-Dunne, E. Mendelowitz, J. Tucek, and R. Pless. Blending participatory culture and urban ecology: Experiments in collaborative imaging for urban forest monitoring. In *18th International Symposium on Electronic Art*, 2012.

[138] A. Zoss and K. Börner. Mapping interactions within the evolving science of science and innovation policy community. *Scientometrics*, 91(2):631–644, 2012.

[139] A. Zoss, M. Conover, and K. Börner. Where are the academic jobs? Interactive exploration of job advertisements in geospatial and topical space. In S.-K. Chai, J. Salerno, and P. Mabry, editors, *Advances in Social Computing*, volume 6007 of *Lecture Notes in Computer Science*, pages 238–247. Springer Berlin Heidelberg, 2010.

8 stuff that i cut out (stephen)

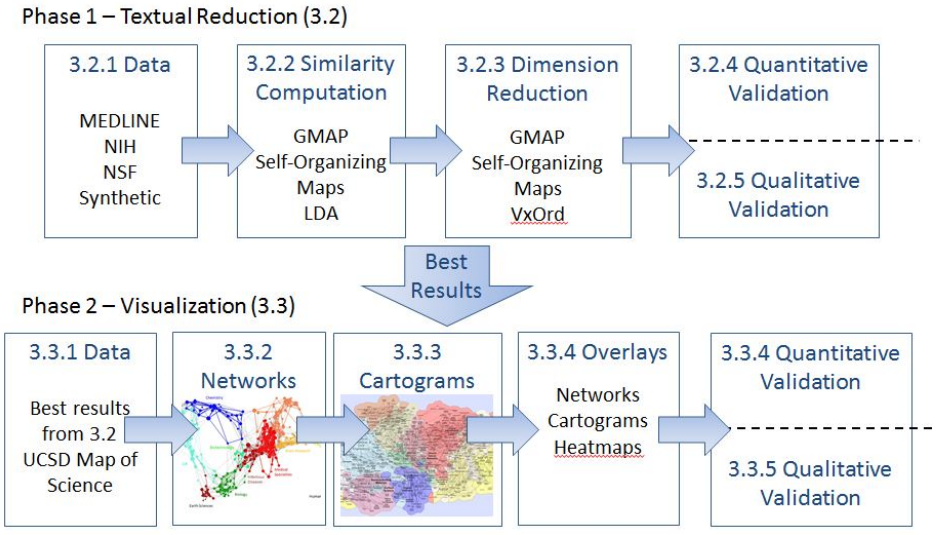

Figure 9: Phase 1 and 2 workflows [[TO BE REPLACED]]

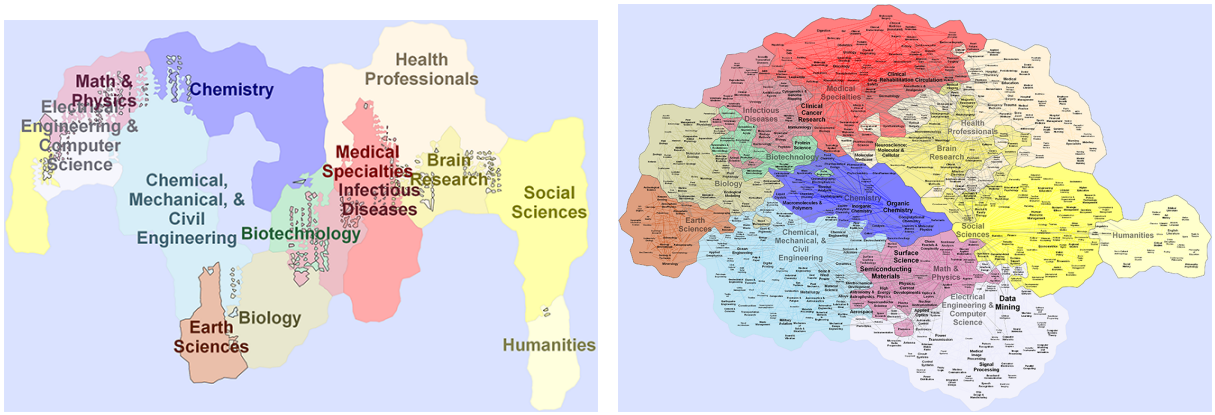

Figure 10: Document density (left) and cluster shapes vs. background (right) [[TO BE FIXED SOMEHOW]]