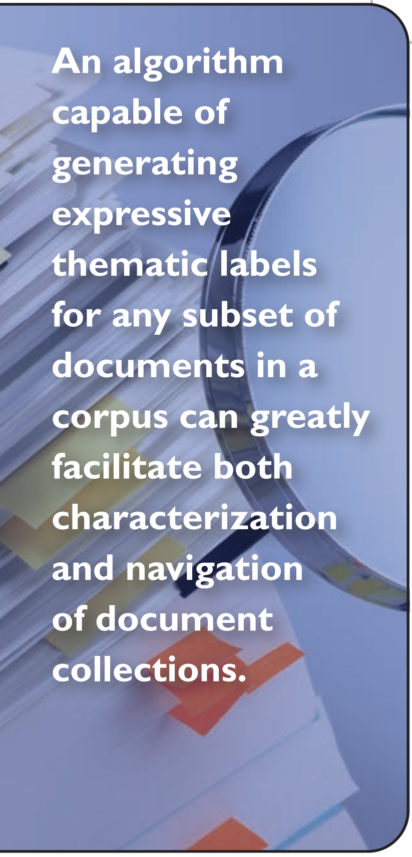
Visual Analytics for Large Document Sets

Arun S. Maiya and Robert M. Rolfe

The Problem

We examine what we refer to as topic similarity networks: graphs in which nodes represent latent topics in text collections and links represent similarity among topics. Efficient and effective approaches to both building and labeling such networks are described. Visualizations of topic models based on these networks are shown to be a powerful means of exploring, characterizing, and summarizing large collections of unstructured text documents



An algorithm capable of generating expressive thematic labels for any subset of documents in a corpus can greatly facilitate both characterization and navigation of document collections.

In our article, we examine network visualizations as a means of enhancing the interpretability of probabilistic topic models for insight discovery. We focus on what is perhaps the most popular and prevalently used topic model: latent Dirichlet allocation or LDA (Blei, Ng and Jordan 2003). Topic modeling algorithms like LDA discover latent themes (i.e., topics) in document collections and represent documents as a combination of these themes. Thus, they are critical tools for exploring text data across many domains. It is often the case that users must discover the subject matter buried within large and unfamiliar document sets (e.g., sensemaking in text data). Keyword searches are inadequate here, since even to begin searching is unclear. Topic discovery techniques such as LDA are a boon to users in such scenarios, because they reveal the content in an unsupervised and automated fashion. However, obtaining a “big picture” view of the larger trends in a document collection from only the raw output of an LDA model can be challenging. In our article, we investigate, the use of what we refer to as topic similarity networks to address this challenge. Topic similarity networks are graphs in which nodes represent latent topics in text collections, and links represent similarity among topics. We described efficient and effective methods to both building and labeling such networks.

Preliminaries

Let $D=\{d_1,d_2,....,d_N\}$ represent a document collection of interest and let K be the number of topics or themes in D. Each document is composed of a sequence of words: $d_i=\langle w_{i1},w_{i2},\ldots,w_{iN_i}\rangle$ where N_i is the number of words in d_i and $i\in\{1\ldots N\}$. Let $W=_{i=1}^{N}f(d_i)$ be the vocabulary of D, where $f(\cdot)$ takes a sequence of elements and returns a set. Probabilistic topic models like LDA take D and K as input and produce two matrices as output. The matrix $\theta\in R^{N\times K}$ is the document-topic distribution matrix, which shows the distribution of topics within each document. The matrix $\beta \in R^{K \times |W|}$ is the topic-word distribution matrix, which shows the distribution of words in each topic. Each row of these matrices represents a probability distribution. For any topic $i \in \{1,\dots,K\}$, the L terms with the highest probability in distribution β_i are typically used as thematic labels for the topic. We use these LDA-derived labels as a baseline for comparison in our work. We begin by describing the construction of the topic similarity network.

Constructing the Network

LDA captures the degree to which both documents and words are topically related. However, relations among the topics themselves are not explicitly captured. In this section, we define these relations by measuring topic similarity.

Measuring Topic Similarity

Recall that topics are represented as probability distributions over vocabulary W and captured by the matrix β. Thus, the similarity for any two topics can be directly computed by comparing the word distributions from β. We employ the Hellinger distance metric to compute topic similarity. Specifically, for any two topics $x,y \in \{1 \dots K\}$, the Hellinger similarity is measured as:

$$H_S(\beta_x,\beta_y)=1-\frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{|W|}(\sqrt{\beta_{xi}}-\sqrt{\beta_{yi}})^2}. \qquad (1)$$

A topic similarity network $G=(V,E)$ can be constructed where $V=\{v_1 \dots v_K\}$ is the set of nodes representing discovered topics and E is the set of edges representing similarities among topics. For any two topics $x,y \in \{1 \dots K\}$, an edge $\{v_x,v_y\} \in V$ exists if and only if $H_S(\beta_x,\beta_y)$ is greater than some predefined threshold, ξ. A MapReduce implementation of these computations is also possible.

Discovering Larger Themes

We employ the use of a community detection algorithm to discover insights into how topics are related to each other and form larger themes. A *community* can be loosely defined as a set of nodes more densely connected among themselves than to other nodes in the network (Blondel, et al. 2008). Within the context of a topic similarity network, such communities should represent groups of highly related topics, which we refer to as topic groups. To detect these communities (or topic groups), we employ the use of the Louvain algorithm, a heuristic method based on modularity optimization (Blondel, et al. 2008).

Labeling the Network

An algorithm capable of generating expressive thematic labels for any subset of documents in a corpus can greatly facilitate both characterization and navigation of document collections. Here, we employ such an algorithm to label nodes in a topic similarity network, as each node is a topic comprising a subset of documents in the corpus. Our approach, referred to as DOCSETLABELER, is a purely unsupervised, extractive method and shown in Algorithm 1. DOCSETLABELER takes D_S, a subset of corpus D, as input, where D_S consists

Algorithm 1 DocSetLabeler algorithm

```
Require: $D_S \subset D$, a subset of corpus $D$
Require: $C$, the number of candidate terms to consider
Require: $L$, the number of labels to return for document set ($L \leq C$)
Require: stopwords, list of terms to filter out
pos = a hash table
neg = a hash table
for all $d \in D$ do
    terms1 = extractSignificantPhrases(d, stopwords)
    terms2 = extractNounPhrases(d, stopwords)
    terms3 = extractProperNounUnigrams(d, stopwords)
    candidates = (terms1 ∩ terms2 ) ∪ terms3
    for all $c \in candidates$ do
        $x$ = normalized frequency of term c in $d$
        $y=1-\frac{\text{index of first occurrence of c in } d}{\text{num. of words in d}}$
        $(\text{weight of term } c) = \frac{2\cdot x\cdot y}{x+y}$
    end for
    If $d \in D_S$ then
        pos[d] = top $C$ terms based on weight
    else
        neg[d] = top $C$ terms based on weight
    end if
end for
for all $\ell \in \bigcup_{x \in pos.values()} X$
    # compute information gain for each label $\ell$
    (score of label $\ell$) = calcScore($\ell$, pos, neg)
end for
top_candidates= top $C$ labels based on information gain
# optionally re-sort final top candidates
top candidates = re_sort(top_candidates)
return top $L$ labels from top_candidates
```

of all documents associated with some LDA-discovered topic $t \in \{1\ldots K\}$. In the present work, D_S is constructed by transforming topics into mutually exclusive clusters, where the topic cluster for document d_i is argmax$_x \theta_{ix}$ (for $i \in \{1\ldots N\}$). Each cluster is an input D_S to Algorithm 1.

DOCSETLABELER is essentially a descriptive model of topic labeling that follows naturally from four observed characteristics of high-quality, topic-representative labels: Expressivity, Prominence, Prevalence, and Discriminability.

Expressivity

Expressivity captures the extent to which labels express and represent themes. Human-assigned labels tend toward multi-word noun phrases, as they are more expressive than unigrams. Unigrams tend to be most expressive when denoting uniqueness (i.e., a proper noun). This is especially true of research reports, our domain of interest, as proper noun unigrams denote important concepts, systems, techniques, or programs (e.g., “LinearSVM,” “F-22”). Lines 4-6 in Algorithm 1 explicitly extract terms conforming to the above principles. The extractSignificantPhrases(•) function uses likelihood ratio tests to extract phrases of multiple words that occur together more often than chance. For a bigram of words w_1 and w_2, this association, *assoc*(·,·), is measured as:

$$assoc(w_1,w_2)=2\sum_{ij} n_{ij}\log\frac{n_{ij}}{m_{ij}}, \qquad (2)$$

where n_{ij} are the observed frequencies of the bigram from the contingency table for w_1 and w_2 and m_{ij} are the expected frequencies assuming that the bigram is independent.

Prominence

Prominence captures the degree to which labels are featured prominently within individual documents. Intuitively, prominent terms tend to make their first appearance earlier and also appear more frequently. Thus, we weight candidate labels by both frequency and position using the harmonic mean, as shown in Line 11 of Algorithm 1.

Prevalence and Discriminability

Good labels for a particular topic appear in many documents pertaining to that topic (*Prevalence*) and appear rarely in other unrelated topics (*Discriminability*). The concept of information gain from the field of information theory simultaneously captures both prevalence and discriminability. Consider a document collection D where documents belong to either a positive or negative category. The *entropy* H of D measures impurity as follows: $H(D)=-p^{+}\log_{2}(p^{+})$-$p$-$\log_{2}$(p-), where p+ and p- are the proportions of positive and negative documents in D, respectively.[1] In Algorithm 1, we assign D_S as positive and $\overline{D_S}$ as negative. The information gain IG of a candidate label ℓ in D, then, is the expected entropy reduction due to segmenting on ℓ:

$$IG(\ell,D)=H(D)-\left(\frac{|D^{\ell}|}{|D|}H(D^{\ell})+\frac{|\overline{D^{\ell}}|}{|D|}H(\overline{D^{\ell}})\right),$$

where D^{ℓ} is the set of documents in D from which label ℓ was extracted. Information gain is computed by the *calcScore*(•) function in Algorithm 1. At the end of the previous step, we are left with a small number of candidate labels (e.g., C=5) for each topic. One can simply select the label with the highest information gain (i.e., the existing sorting) or re-sort based on a combination of other factors (e.g., label frequency, word probabilities from β), as indicated in Line 25 of Algorithm 1.

[1] Note that $\log^{2}(0)$ is taken to be 0.

Case Study: NSF Research Grants

As a realistic and informative case study, we utilize our methods to characterize and visualize basic research funded by the National Science Foundation (NSF). The corpus considered in this case study consists of 132,372 titles and abstracts describing NSF awards for basic research between the years 1989 and 2003 (Bache and Lichman 2013). We executed the MALLET implementation of LDA (McCallum 2002) on this corpus using K=400 as the number of topics and 200 as the number of iterations. All other parameters were left as defaults. For topic similarity, we experimentally set ξ as 0.15 to yield a graph density of approximately 0.01. For the labeling of topic nodes in the network using DOCSETLABELER, we set C=5 and L=1.

Topic Labeling of NSF Grants

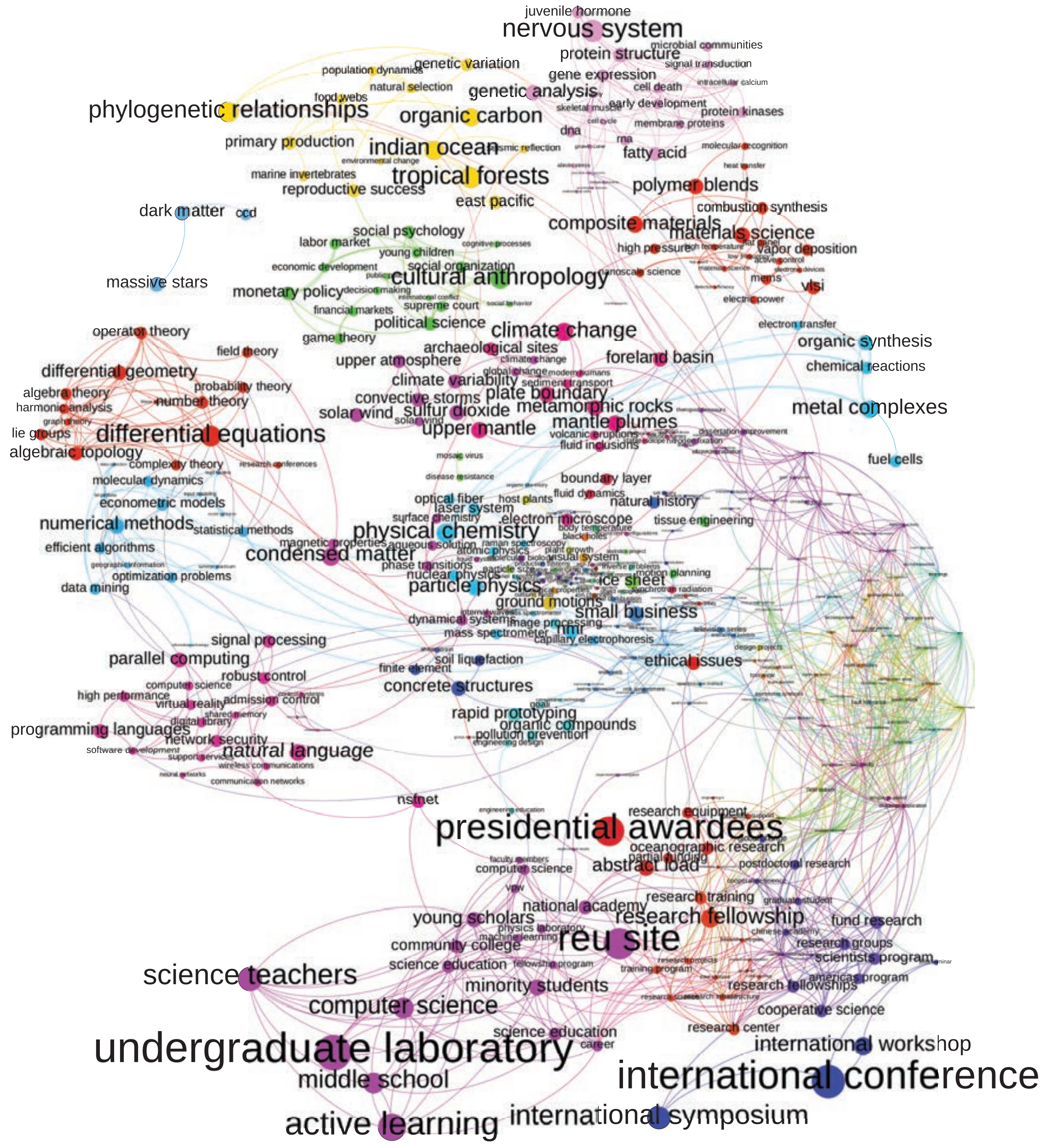
Table 1 shows the labels generated for a sample of ten discovered topics by both DOCSETLABELER and LDA. Labels produced by DOCSETLABELER are more expressive and representative of the true themes of each topic. We assigned two judges to evaluate labels for all topics. For a fair comparison, we showed six unigram labels from LDA but only three labels (mostly bigrams) from DOCSETLABELER for each topic. As shown in Table 2, both judged the labels by DOCSETLABELER to be generally superior (χ^2=145.73, P<0.0001) with an inter-judge agreement of 0.62, as measured by Cohen's kappa coefficient.

Visualizing NSF Grants

A topic similarity network was constructed, with each node representing a topic and labeled using the highest ranked term returned by DOCSETLABELER. The network, which concisely presents a comprehensive and holistic view of roughly 15 years of NSF-funded research, can be navigated and explored using any available network visualization software (e.g., Gephi, Cytoscape). The entire network is shown in Figure 1, where both expected and unexpected trends are revealed. The visualization encapsulates the major research funding efforts for scientific research in addition to the subtle connections among them. Major funding efforts for education and conference support are also displayed (toward the bottom). In this network and all networks shown in our article, node sizes indicate the number of documents pertaining to the topic represented by the node. Sizing nodes by funding amount is also possible. Node colors indicate the community (or topic group) affiliation. Using this network, one can better understand how topics form larger themes, discover and characterize information of interest, and derive insights into how best to search and explore the corpus further. We present illustrative examples of the patterns and trends discovered using our topic similarity network. Figure 2 shows one small corner of the "topic universe" — a "social clique" of math topics discovered by community detection within the larger network of all topics. Note that each node in the network represents hundreds of documents (or more). Thus, this visualization of math topics clearly

Table 1. [NSF Grants.] Ten discovered NSF topics and the highest-ranked labels assigned to each by both LDA and DocSetLabeler.

Actual Topic	Labels from LDA	Labels from DocSetLabeler
Fluid Mechanics and Fluid Dynamics	Flow, fluid, flows, fluids, dynamics, transports	Fluid dynamics, fluid mechanics, multiphase flow
Game Theory	Agents, theory, game, agent, games, equilibrium	Game theory, economic agents, repeated games
Graph Theory	Discrete, graph, combinatorial, theory, combinations, graphs	Graph theory, algebraic combinatorics, ramsey theory
Human Evolution	Modern, fossil, early, years, human, age	Modern humans, human evolution, hominid evolution
Hydrology	Water, river, hydrologic, watershed, balance, surface	Hydrologic controls, watershed scale, alpine basins
Modal Analysis in Structural Engineering	Mode, modes, research, vibration, direction, coupling	Normal modes, vibration control, modal analysis
Object Recognition	object, objects, features, recognition, oriented, feature	Object recognition, curved objects, cluttered scenes
Protein Function/Mechanisms	Protein, proteins, function, role, biochemical, phosphorylation	Protein kinases, protein phosphorylation, protein import
Protein Struction	Protein, proteins, binding, structure, amino, acid	Protein structure, protein folding, amino acid
Social Psychology	Social, people, research, individuals, attitudes, status	Social psychology, social influence, social perception



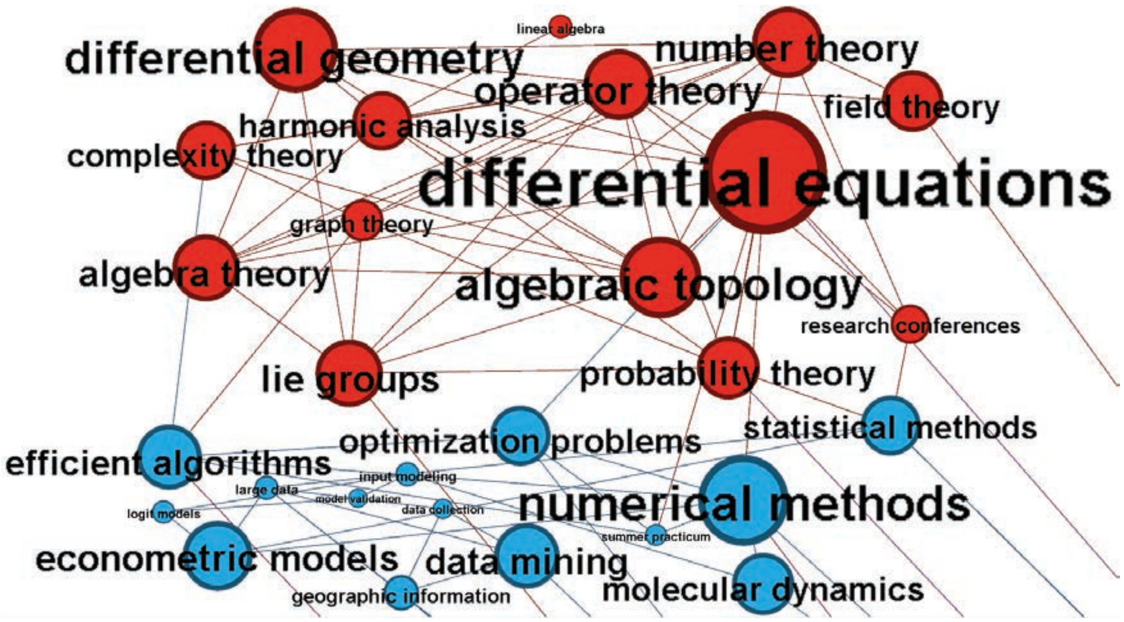
Major research topics – including their subtle connections to each other – are shown. Also displayed (toward the bottom of network) are major funding efforts for education support and conference support. Node sizes indicate the number of grant abstracts pertaining to the topic. Node colors indicate the community (or topic group) affiliation, which illustrate how research topics form larger themes.

Figure 1. [NSF Grants] Topic Similarity Network of Roughly 15 years of NSF research and support (i.e., a Total of 132,372 Research Grants)

Table 2. Evaluation of Labels for Each Method on NSF Grants

	DocSetLabeler	LDA
DocSetLabeler	313	6
LDA	23	29

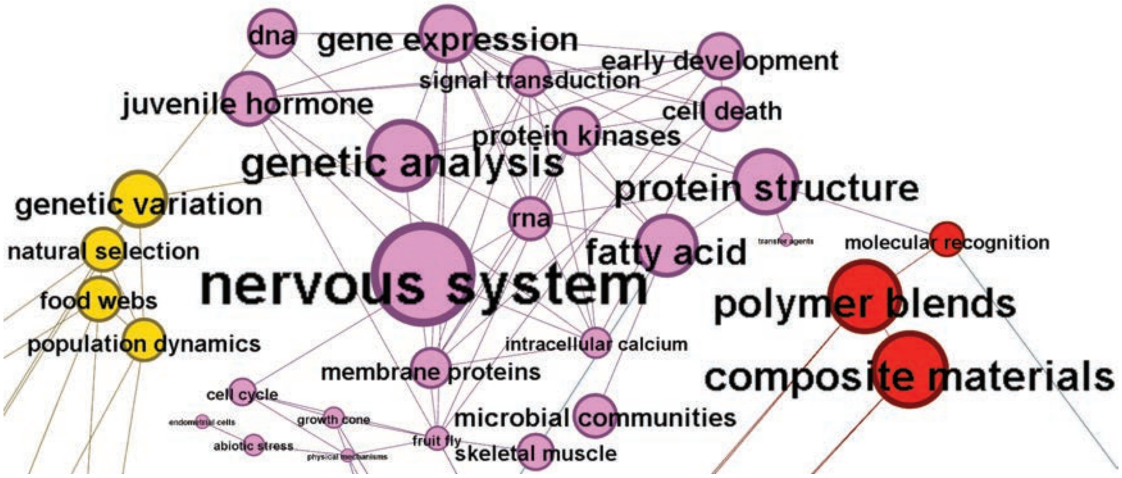
Overall, both judges chose labels from DOCSETLABELERbe most on-point.



The red color covers pure math, while the blue is more applied.

Figure 2. [NSF Grants] Two Discovered Topic Groups (or Communities) Pertaining to Math-Oriented Research

and concisely summarizes more than 10,000 documents. Such visualizations also provide insights into relations between topic groups. For instance, Figure 3 shows a community of biology-related topics (shown in pink). Here, we see peripheral connections to another life science theme (shown in yellow) containing topics such as *genetic variation, population dynamics*, and *food webs*. We also see a peripheral connection to a material science theme (shown in red), illuminating research areas dedicated to developing materials based on biological and organic components and also the mutual interest in molecular recognition. As a final example, Figure 4 shows a connected component of astronomical research topics that appears separate from the larger network. This last example



The topic group related to biology is shown in pink. Also shown are topic nodes from other related communities (e.g., polymer blends, population dynamics) and their peripheral connections to this biology-related topic group.

Figure 3. [NSF Grants] A Discovered Topic Group Related to Biology (shown in pink)

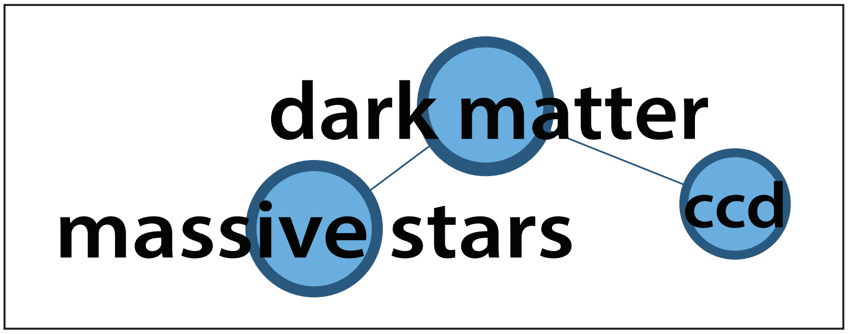


Figure 4. [NSF Grants.] Connected Component of Astronomical Research Topics Separated from the Larger Network

illustrates one possible way to use these visualizations to identify outliers (i.e., topics that are comparatively more different than the larger corpus based on their set of similarity scores).

For additional results and technical details for this analysis, refer to the full report (Maiya and Rolfe October 2014).

***Dr. Maiya** is a Research Staff Member in IDA's Information Technology and Systems Division. He holds a Doctor of Philosophy in computer science from the University of Illinois at Chicago.*

***Dr. Rolfe** is a Research Staff Member in IDA's Information Technology and Systems Division. He holds a Doctor of Philosophy in physics from the University of California, Los Angeles.*

The original article was published in *Proceedings of the 2014 Institute for Electrical and Electronics Engineers (IEEE) International Conference on Big Data.*

"Topic Similarity Networks: Visual Analytics for Large Document Sets"

http://dx.doi.org/10.1109/BigData.2014.7004253

References

Bache, K., and M. Lichman. 2013. "UCI machine learning repository."

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." J. Mach. Learn. Res. 3 (4-5): 993-1022.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast unfolding of communities in large networks." Journal of Statistical Mechanics: Theory and Experiment 10: P10008+.

Maiya, Arun S., and Robert M. Rolfe. October 2014. "Topic similarity networks: Visual analytics for large document sets." IEEE International Conference on Big Data (Big Data). 364-372.

Maiya, Arun S., John P. Thompson, Francisco L. Loaiza-Lemos, and Robert M. Rolfe. 2013. "Exploratory Analysis of Highly Heterogeneous Document Collections." Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13. New York, NY: ACM. 1375–1383.

McCallum, Andrew K. 2002. "MALLET: A Machine Learning for Language Toolkit."