Abstract—Document similarity graphs are a useful visual metaphor for assessing the conceptual content of a corpus. Algorithms such as Latent Dirichlet Allocation (LDA) provide a means for constructing such graphs by extracting topics and their associated term lists, which can be converted into similarity measures. Given that users’ understanding of the corpus content (and therefore their decision-making) depends upon the outputs provided by LDA as well as how those outputs are translated into a visual representation, an examination of how the LDA algorithm behaves and an understanding of the impact of this behavior on the final visualization is critical. We examine some puzzling relationships between documents with seemingly disparate topics that are linked in LDA graphs. We use TopicView, a visual analytics tool, to uncover the source of these unexpected connections.

Index Terms—Visual analytics, Text analysis, Latent Dirichlet Allocation.

1 INTRODUCTION

In working with similarity graphs generated from Latent Dirichlet Allocation (LDA) [1] models, we were struck by the high degree of connectivity in the graph. Many of the edges were unexpected, connecting documents covering seemingly unrelated topics. Why was LDA linking these documents? We felt that understanding the cause of these links was essential, since their presence alters both the graph layout and its interpretation by users [5].

To explore this problem, we used visual analytics to reveal how the LDA algorithm makes the connections shown in a similarity graph. Our tool TopicView [2] combines a user-level view of the similarity graph with linked views that enable exploration of the relationships between documents, topics, and terms. This paper describes our investigation’s visual analysis and the improved separation of topic groups we were able to achieve as a result.

2 BACKGROUND OVERVIEW

This section provides a brief description of the data set that we used in our analysis, the LDA algorithm and associated parameter values, and the TopicView visualizations used.

2.1 Data Set

For this analysis, we used the DUC data set, a collection of newswire documents from the Associated Press and New York Times that were used in the 2003 Document Understanding Conference (DUC) to evaluate document summarization systems [4]. The collection contains 30 clusters comprising 298 documents. Each cluster contains roughly 10 documents focused on a particular topic or event. The documents all have human-generated cluster labels, which we use to color-code documents. We informally define a cluster as a group of documents exhibiting strong links between members within the group and weak links outside the group. As with any non-trivial, real-world corpus, the topics in each cluster are not entirely disjoint in term distributions; rather, each topic focuses upon a separate news event. Thus, there are some conceptual commonalities that reasonably can be considered as possible linking mechanisms between clusters.
2.2 Latent Dirichlet Allocation

LDA is a hierarchical probabilistic generative model used to model a collection of documents by topics, i.e., probability distributions over a vocabulary [1]. Given a vocabulary of \( W \) distinct words, a number of topics \( K \), two smoothing parameters \( \alpha \) and \( \beta \), and a prior distribution (typically Poisson) over document lengths, this generative model creates random documents whose contents are a mixture of topics.

In order to use LDA to model the topics in an existing corpus, the parameters of the generative model must be learned from the data. Specifically, for a corpus containing \( D \) documents we want to learn \( \phi \), the \( K \times W \) matrix of topics, and \( \theta \), the \( D \times K \) matrix of topic weights for each document. The remaining parameters \( \alpha \), \( \beta \), and \( K \) are specified by the user. For the LDA models used in this paper, parameter fitting is performed using collapsed Gibbs sampling [3] to estimate \( \theta \) and \( \phi \).

We set \( K = 30 \) to match the number of anticipated clusters in the corpus. Following Blei et al. [1], we use \( \alpha = 50/K \) and \( \beta = 0.1 \). Two additional parameters for the Gibbs sampling are the number of sampling and burn-in iterations, which we set to 1 and 200, respectively.

2.3 TopicView

Although TopicView was developed to compare LSA models with LDA models [2], for this paper we use it to explore individual LDA models. We predominantly use the following four views: the Term Table, the Document Similarity Graph, the Document-Topic Table, and Document Text. Documents can be selected either through the graph of document relationships or the table of document model features, highlighting the selection in both views and displaying the selected document contents in the Document Text view.

2.3.1 Term Table

The Term Table shown in Figure 2 presents the terms for each topic (i.e., the rows of the topic matrix \( \phi \)) sorted in decreasing order of importance. Text color provides an additional cue about the relative weights of terms. Terms with the highest weights are drawn in black, fading to gray for the lowest-weighted terms. Since we are most interested in distinguishing weighting differences at the high end of the scale, we spread this part of the range by using a logarithmic mapping that increases the number of luminance steps as we approach black. Individual terms are selectable. Once selected, each instance of that term within every topic is highlighted with a lighter background. The selection is linked to the Document Text view, where every instance of that term within the selected documents is highlighted in red.

2.3.2 Document Similarity Graphs

In TopicView, we compute edge weights between every pair of documents by calculating the cosine similarities of the topic weight matrix \( \theta \). To reduce visual clutter, we threshold edges by keeping the strongest links, while retaining a minimum number of edges per node. We determine which edges to keep on a document-by-document basis as follows: (i) sort each document’s edges in descending order by weight, (ii) keep all edges with weights greater than a significance threshold (we use 0.9), and (iii) if a document’s edge count is less than a minimum (5 in our examples), add edges in diminishing weight order until it is reached.

We project the graphs into two dimensions using a linear time force-directed layout algorithm. Each document node is labeled with its document ID and color-coded using the ground-truth category from DUC. Edges are color-coded using saturation to indicate similarity weights, with low values in gray and high values in red. We highlight selected nodes in white, which is the one color not used for the DUC label categories. Selected edges are in blue.

2.3.3 Document-Topic Table

The Document-Topic Table shown in Figure 3 is LDA’s \( \theta \) matrix of topic weights for each document. In a manner identical to the Term Table, the values in the table are varied between black and light gray to permit rapid visual scanning of rows and columns for darker, more highly weighted documents within a topic. Although there is a tendency to try to identify topics with clusters, the weightings shown in the Document Table demonstrate that document groups frequently contribute in varying degrees to multiple topics (weightings spread across rows). Similarly, topics typically include multiple document groups (weightings spread across columns).

The visible columns within the table are controlled by selecting topics. Subsetting the column display facilitates side-by-side comparisons of the relative weightings of the most significant documents associated with a set of topics. The table can be used to formulate hypotheses about the relationship between conceptual content and specific documents.

3 Analysis

We start by examining documents with seemingly unrelated topics, such as document 3 shown in the center of the graph in Figure 2. According to the text of documents 3 and 4, both are stories about Pinochet’s arrest in Britain. Document 96 is a story about a Yugoslavian tribunal to prosecute Bosnian war crimes. Finally, document 193 is about cold weather killing 39 people in Moscow. We project the graphs into two dimensions using a linear time force-directed layout algorithm. Each document node is labeled with its document ID and color-coded using the ground-truth category from DUC. Edges are color-coded using saturation to indicate similarity weights, with low values in gray and high values in red. We highlight selected nodes in white, which is the one color not used for the DUC label categories. Selected edges are in blue.

2.3.3 Document-Topic Table

The Document-Topic Table shown in Figure 3 is LDA’s \( \theta \) matrix of topic weights for each document. In a manner identical to the Term Table, the values in the table are varied between black and light gray to permit rapid visual scanning of rows and columns for darker, more highly weighted documents within a topic. Although there is a tendency to try to identify topics with clusters, the weightings shown in the Document Table demonstrate that document groups frequently contribute in varying degrees to multiple topics (weightings spread across rows). Similarly, topics typically include multiple document groups (weightings spread across columns).

The visible columns within the table are controlled by selecting topics. Subsetting the column display facilitates side-by-side comparisons of the relative weightings of the most significant documents associated with a set of topics. The table can be used to formulate hypotheses about the relationship between conceptual content and specific documents.
Fig. 3: The document-topic weights for document 3 (top image) and its selected neighbors.

Fig. 4: All of the bridging documents are in white with the connecting edges in blue.

discuss additional terms such as Iran in topic 36 later.) The selection of these four topics comes from an examination of the document-topic weights for each document shown in Figure 3. We have included all topic columns that have darker/higher weights within the selected rows, including topic 30, whose most significant terms do not match any of the story lines seen in the documents’ texts. Examining the terms in topic 30, we find that XML document tags (“headline”, “slug”, “body”, etc.) predominate. As shown in the left image in Figure 1, topics 30 and 31 capture Associated Press (AP) and New York Times (NYT) document origins, respectively. The connection between these seemingly dissimilar documents is that they are all AP articles.

3.1 Bridging Documents

Looking at the weights in the Document-Topic Table images in Figure 3, an interesting pattern emerges. Documents 0–9 (top image) are articles about Pinochet’s arrest. The human-generated cluster labels (brown color-coding in the document ID column) show that these documents all belong to this group. LDA has similarly identified this same group of documents as a group, shown by the darker text of the stronger weights in column 44. The weights for documents 0 through 4 in column 30 show a significant connection between these documents and the topic, articles from AP, whereas documents 5 to 9 do not. Unlike document 4, the weighting for document 3 in topic 30 is stronger than its weighting in topic 44 (i.e., document 3 is more strongly aligned with its AP source than with its conceptual content). This same pattern is seen with documents 96, 142, 144, and 193. We hypothesize that documents matching this pattern are the source of many of our edges between disjoint topics.

We test our hypothesis by listing all documents whose strongest weights are within the AP or NYT topic columns (30 and 31), then we check the conceptual content of these documents against all documents directly linked to them in the Document Similarity Graph. If the linked documents can be seen as having a connection in terms of their content, then the list document fails the hypothesis and is removed. We define a conceptual connection to exist between a list and linked document if both are strongly weighted in the same topic column in the Document-Topic Table, if a number of common terms are found in their document texts, or if their human-labeled categories match.

Of the original 297 documents, 33 fit our hypothesis. Of those, only 11 survived our test and exhibited links that we could not account for in some other way, including the ones originally observed (3, 96, 142, 144, and 193). The 11 documents and the edges connecting them are shown in Figure 4. These documents are in the center of the graph and all of them tend to link with one another, impacting the layout of their associated clusters and creating so many edge crossings that the true connectivity is difficult to follow. The full graph is shown in Figure 1.

3.2 Tag Terms

All of the bridging documents are AP articles. All are short, with the story content sometimes being only a single sentence. Comparing
the human-labeled cluster boundaries. Once again, seemingly unre-
really an election story, which validates its connection to the Iranian
highly ranked terms, and they appear in document 142 three and four

to the Iranian election topic, 40, than the Kosovo tribunal topic. Docu-
ments 19, 10, 144, and 265 (found near the center of the graph in
Figure 5) here, in a manner similar to the AP topic, the individ-
ual documents are more strongly connected to each other through this
subtopic than they are to their cluster groups. This topic acts as a
bridge connecting all of the clusters together.
Clearly, our initial choice of 30 topics is impacting the resulting
clusters, but given LDA’s merging and splitting of topics, it is difficult
to select an appropriate value. Experimenting with various topic val-
ules between 28 and 75, we find that increasing the topic count does not
necessarily separate combined clusters or reduce the number of edges.
We find new bridging topics, and topics that combine into new merged
clusters. In addition, as the number of topics expands, the document
weightings for a subset of the additional topics becomes so low as to
make the topics appear to be noise.

4 Conclusions
LDA’s choice of topics may include unexpected categories, such as ar-
ticle sources, if the document text contains term distributions that can
act as signatures for those sources. For example, using LDA to model
VisWeek’s digital proceedings generates a topic consisting of all the
HTTP references. In turn, these additional topics act as bridges be-
tween conceptual topics, linking seemingly unrelated articles. Short
documents facilitate this by being more strongly connected to their
source topic than to their conceptual content. Although we were able
to remove source-specific terms from our documents and generate only
thematic topics, this only works if source-based and concept-based ter-
mology differ. If the various sources have unique writing styles that
rely on a distinct vocabulary, this filtering may not be possible. The
issues around selecting the best number of topics and counteracting
the tendency of short documents to act as bridges between dissimilar
document groups remain to be solved in future work.

Whether this bridging is seen as having a positive or negative im-
 pact depends upon the application. If the user is trying to understand
just the thematic content of a corpus, additional document connections
blur the thematic boundaries and, in the worst case, obscure the very
patterns the user is hoping to see. However, if the user is trying to
connect the dots between disparate bits of information, where sources
provide important clues or impact the reliability of the answer, source
connections may be desirable. Understanding these subtleties allows
application designers to make conscious choices about the combined
impact of the analysis and the visual representation on the users’ un-
derstanding of the data.

Acknowledgments
This work was funded by the Laboratory Directed Research & De-
velopment (LDRD) program at Sandia National Laboratories. Sandia
National Laboratories is a multi-program laboratory managed and op-
erated by Sandia Corporation, a wholly owned subsidiary of Lockheed
Martin Corporation, for the U.S. Department of Energy’s National Nu-
clear Security Administration under contract DE-AC04-94AL85000.

References
Visually comparing topic models of text collections. In Proceedings 23rd
IEEE International Conference on Tools with Artificial Intelligence, 2011.
generic news text summarization systems. In Proc. DUC 2003 workshop
on text summarization, 2003.
to conceptual similarity. IEEE Transactions on Visualization and Com-