Topicalizer: Reframing Core Concepts in Machine Learning Visualization by Co-designing for Interpretivist Scholarship

ABSTRACT

Computational algorithms can provide novel, compelling functionality for interactive systems. However, designing such systems for users whose expertise lies outside computer science poses novel and complex challenges. This paper focuses specifically on the domain of designing interactive topic modeling visualizations to support interpretivist scholars. It describes a co-design process that involved working directly with two such scholars across two different corpora. The resultant visualization has both several similarities and key differences with other topic modeling visualizations, illustrated here using both the final design and discarded prototypes. The paper's core contribution is an attention to how our emphasis on interpretation played out, both in the design process and in the final visualization design. The paper concludes by discussing the kinds of issues and tensions that emerged in the course of this work, as well as the ways that these issues and tensions can apply to much broader contexts of designing interactive algorithmic systems.

CONTENTS

1. INTRODUCTION

Designing computational systems means, increasingly, designing for interactions with algorithms. Machine learning, recommender systems, sentiment analysis, and other algorithmic techniques have become common in interactive, user-facing systems. These developments raise novel, complex challenges, both in the design of such systems (Baumer, 2017; Dove et al., 2017; Leahu, 2016; Yang et al., 2018) and in terms of users' interactions with them (Ananny, 2011; Bucher, 2017; Eslami et al., 2015, 2016; Gillespie, 2012, 2013).

Computational tools for supporting interpretivist text analysis by sociological and humanistic scholars provides a compelling context in which to explore these issues. Given a large corpus, how does one determine what the documents are about? Traditionally, methods to address this question include qualitative analysis (Lofland et al., 1971 (2005)), grounded theory (Charmaz, 2006; Glaser & Strauss, 1967), discourse analysis (Foucault, 1972), and others. However, it can be difficult to analyze copious volumes of unstructured text via such "close reading" methods, which are time- and labor-intensive.

One increasingly common approach leverages computational techniques in areas such as digital humanities (Underwood, 2019; Jockers, 2013; Jockers & Mimno, 2013; Rhody, 2013; Underwood, 2014) or computational social science (Lazer et al., 2009; Roberts et al., 2014). Researchers employ sentiment analysis (Pang & Lee, 2008), topic modeling (Blei et al., 2003; Blei, 2012), collocate analysis (Lind & Salo, 2002), and a variety of other techniques to analyze large volumes of text.

Furthermore, computational methods do more than simply scale up human ability. Humans and computers read texts in different ways (Taylor, 2009; Burrell, 2016; Passi & Jackson, 2017). Computational systems provide an alternative lens, a different view of textual data that might not have been achieved by human researcher(s) alone (Baumer et al., 2017; Muller et al., 2016). Put differently, computational analysis is not simply faster but rather a fundamentally different means of text analysis.

These developments create both exciting possibilities and significant challenges. While some social scientists and humanists are cross-training in the use of computational tools, it is unrealistic to assume that all will become experts. It may be more productive, for instance, if a sociologist or humanist spends their time reading, analyzing, and interpreting text data rather than debugging R code.

This line of reasoning points to an important design space: How can we make the results of computational text analyses approachable for and interpretable by researchers whose expertise lies outside computing? Prior work has explored different means of visualizing computational analyses such as topic modeling (Chaney & Blei, 2012; Choo et al., 2013; Chuang, Manning, & Heer, 2012; Chuang, Ramage, et al., 2012; Dinakar et al., 2015; Goldstone, n.d.; H. Lee et al., 2012; Mimno, 2013). However, most such work is driven more by the underlying mathematical formalisms than by the kinds of research questions various scholars are seeking to address using topic modeling (for an exception, see Klein & Eisenstein, 2013; Klein et al., 2015).

In contrast, this paper presents both the process and product of *co-designing a topic modeling visualization for interpretive text analysis*. The design process is described via an analogy to the train-test paradigm in machine learning (Gelman & Hill, 2006). Initially, the visualization was iteratively designed and redesigned for sociological analysis of a corpus of blog posts written by parents of children on the autism spectrum. One might refer to this as the "training" phase.

After being developed in this fashion for several months, the visualization was then applied to the personal diaries of a 19th century writer. This step in the process could be referred to a the "testing" phase. Seen through this train-test lens, our approach also resembles case-study based techniques for evaluating information visualizations (Shneiderman & Plaisant, 2006), as well as visualization design study methods (Sedlmair et al., 2012). Both in those approaches and in the work presented here, researchers and designers work closely with domain experts to iteratively redesign the interface (e.g., Perer & Shneiderman, 2008).

That said, this work's emphasis on interpretivist scholarship offers several key differences from prior approaches. Such differences play out not only in the appearance and interaction design of the visualization, but also in the underlying algorithmic techniques for analyzing the text. Throughout, this paper describes design decisions that were made, as well as potential designs that were abandoned, in service of supporting this interpretivist approach. Distilling across the two design cases, the discussion section then summarizes three major themes where an emphasis on interpretation had the most significant impacts: performance evaluation, the constitution of individual topics, and choices to make or to avoid explicit computational and/or visual representations.

Thus, this paper offers a two fold contribution. First, it presents a novel visualization design for topic modeling. The visualization is described both in terms of the final form and in terms of early prototypes that were abandoned, along with the rationales for doing so. The design itself both draws on and expands beyond most prior topic modeling visualization techniques (Chaney & Blei, 2012; Choo et al., 2013; Chuang, Manning, & Heer, 2012; Chuang, Ramage, et al., 2012; Dinakar et al., 2015; Goldstone, n.d.; H. Lee et al., 2012; Mimno, 2013) by emphasizing an interpretivist approach. That is, the visualization is designed less to provide a definitive statement about *the* topics in a corpus and more to support the iterative processes of interpretation and reinterpretation by social scientific and humanist researchers (cf. Baumer et al., 2017; Klein et al., 2015; Muller et al., 2016). Second, this paper describes how designing for interpretation offers alternative approaches to many traditional concepts in machine learning (ML) and artificial intelligence (AI). Across issues related to performance evaluation, data preprocessing, model selection, and meaning making, the paper draws out three important tensions: balancing how well the model performs vs. what the model means, balancing prescription vs. interpretability, and balancing explicit modeling vs. interpretive flexibility. These tensions are brought to the fore here due to the focus on supporting interpretivist scholarship. However, they also resonate with themes arising from the broader literature on the use and interpretation of algorithmic systems (Eslami et al., 2016, 2015; Ananny, 2011; Burrell, 2016; Gillespie, 2012, 2013). Furthermore, navigation of such tensions will likely prove important in future work, not only on supporting interpretivist scholars, but also on designing interactive systems that incorporate AI and ML algorithms more broadly.

2. RELATED WORK

2.1. Topic Modeling Visualization

Prior work has visualized topic modeling results from a variety of corpora. Examples include the Proceedings of the Modern Languages Association (PMLA) (Goldstone, n.d.; Goldstone & Underwood, 2014), the journal Science (Blei & Lafferty, 2007b,a), dissertations completed at Stanford University (Chuang, Ramage, et al., 2012), Wikipedia (Chaney, 2012; Chaney & Blei, 2012), US

Presidential State of the Union addresses (Mimno, 2013), the papers of Thomas Jefferson (Klein & Eisenstein, 2013; Klein et al., 2015), and others.

These systems generally have one of two goals. First, they can be used as an analytic tool. For example, the Wikipedia topic browser (Chaney, 2012; Chaney & Blei, 2012) shows prevalent topics among Wikipedia articles. For documents with a date, jsLDA (Mimno, 2013) shows changes in the prevalence of each topic over time. The Stanford Dissertation Browser (Chuang, Ramage, et al., 2012) allows users to identify similarities and differences among various academic departments at Stanford based on the topics in those departments' dissertations. Fathom (Dinakar et al., 2015) helps crisis hotline counselors who join a conversation already underway quickly understand the caller's situation and emotional state. Second, these systems can be used as a diagnostic tool to inspect and to improve the quality of topic modeling results. TopicCheck (Chuang et al., 2015) shows the stability of individual topics across multiple solutions with different random initializations. UTOPIAN (Choo et al., 2013) allows users to tweak a topic model currently being fitted by manually splitting or combining topics induced by the algorithm. Termite (Chuang, Manning, & Heer, 2012) explores different techniques for labeling and understanding the coherent themes that each topic represents. iVisClustering (H. Lee et al., 2012) allows users to adjust various parameters of a topic model to see how they impact subsequent clustering of documents. A series of experiments by Smith et al. (2017) show how different visual representations of topics can what users perceive a topic to be about and the kinds of labels they generate for topics.

The above work provides a rich variety of visualization techniques on which to draw. However, most of the above designs were guided mostly by the technical constraints of topic modeling. Even in the case of tools designed with human users in mind, such as Fathom (Dinakar et al., 2015), UTOPIAN (Choo et al., 2013), or Termite (Chuang, Manning, & Heer, 2012), the visualization is driven primarily by the underlying mathematical and computational formalisms. Two exceptions to this trend, TOME (Interactive TOpic Model and MEtadata Visualization) (Klein et al., 2015) and Serendip (Alexander et al., 2014), come from work related to digital humanities and are described further below.

The work presented here differs from the above in two important ways. First, it uses a participatory approach (Muller, 2012; Kuhn & Muller, 1993; Ehn, 1988). Rather than focusing on the technical affordances and constraints of topic models, the design is instead driven by the kinds of questions that interpretivist researchers ask. This point connects with work on how people whose expertise lies outside computer science interpret complex algorithmic systems (Eslami et al., 2016, 2015; Ananny, 2011; Bucher, 2017; Gillespie, 2012, 2013). Second, it emphasizes an interpretivist stance (Orlikowski & Baroudi, 1991). Much social scientific research on topic modeling follows a positivist paradigm. Anonymous surveys, controlled experiments, and related techniques are used to ascertain objective facts about social reality (e.g., Roberts et al., 2014). In contrast, the work presented here explores how computational analysis can support interpretivist work. Such work seeks to understand the means by which social groups mutually co-construct and interpret their reality (Berger & Luckmann, 1966; Orlikowski & Baroudi, 1991). This point resonates with prior work on incorporating machine learning into interpretivist-influenced design, particularly in affective computing (Leahu, Schwenk, & Sengers, 2008; Leahu & Sengers, 2014; Leahu, Sengers, & Mateas, 2008; Sengers et al., 2008; Boehner et al., 2007). It also builds on significant work using topic modeling in the humanities and social sciences.

2.2. Interpretivist Computational Text Analysis

A variety of analytic techniques have adapted by computational social science, digital humanities, and related endeavors. One popular approach is topic modeling, e.g., Latent Dirichlet Allocation (LDA) (Blei et al., 2003). This technique identifies latent themes, or "topics," in a corpus of documents without the need for any labeling or training data. Prior work has applied LDA or similar techniques to literature (Jockers, 2013; Jockers & Mimno, 2013), survey data (Roberts et al., 2014), political campaign speeches (Boydstun et al., 2014), poetry (Rhody, 2013, 2017), historical archives (Klein & Eisenstein, 2013; Klein et al., 2015), the scholarship of humanists (Goldstone, n.d.; Goldstone & Underwood, 2014), and others. Boyd-Graber et al. (2017) review how topic modeling has been applied across a variety of domains.

One emerging question deals with the relationship between computational techniques and traditional practices of qualitative (text) analysis (Charmaz, 2006; Corbin & Strauss, n.d.; Glaser & Strauss, 1967; Lofland et al., 1971 (2005)). Recent work has explored different means by which analysis of qualitative data, especially coding (cf. Charmaz, 2006; Glaser & Strauss, 1967; Lofland et al., 1971 (2005)), can be supported by combinations of machine learning, visualization, and other computational techniques (Drouhard et al., 2017; Chen et al., 2018; Zade et al., 2018; Marathe & Toyama, 2018). Some work has argued that, either in theory (Muller et al., 2016) or in practice (Baumer et al., 2017), topic modeling in particular bears some key similarities to qualitative methods such as grounded theory (Charmaz, 2006; Glaser & Strauss, 1967) or open coding (Lofland et al., 1971 (2005)), e.g., both follow a bottom-up approach, and both emphasize iterative refinement. Other work has employed a human-in-the-loop approach to tune topic models dynamically during training (T. Y. Lee et al., 2017; Poursabzi-Sangdeh et al., 2016; H. Lee et al., 2012; Choo et al., 2013; Chuang et al., 2015). Such approaches, it is argued, ensure greater alignment between computational results and human interpretation thereof.

This point about alignment between computational techniques and human research processes applies equally to the design of visualizations. Some work has suggested applying feminist perspectives to visualization (Knigge & Cope, 2006; D'Ignazio & Klein, 2016). Doing so, it is argued, can highlight which epistemological perspectives are dominant and which have been marginalized, as well as support a pluralism of epistemic perspectives (cf. Bardzell, 2010) in visualization. Although they do not incorporate visualization, Concannon et al. (2018) similarly argue for the value of a feminist perspective in interpreting topic modeling about nursing mothers' reviews of different locations for breast feeding.

Such arguments align closely with the approach taken in this paper. To the authors' knowledge, TOME (Klein et al., 2015), and its precursor TopicViz (Eisenstein et al., 2012; Klein & Eisenstein, 2013), offer the only example of a topic modeling visualization specifically designed around interpretivist research processes. That work provides a useful articulation of the issues involved in the design of such visualizations, especially in terms of aligning the design less with the mathematical structure of topic models and more with the human researcher processes the tool is meant to support. However, Klein et al. (2015, i136), in comparing different prototypes, note that "only real user experiences can provide the definitive answer on which design is most useful." In another example, the topic modeling visualization Serendip (Alexander et al., 2014) was used by a literature scholar to examine early modern literature published between 1530 and 1799. Although "the tool was extensively refined based on his [this scholar's] experience" (Alexander et al., 2014, p. 180), little is provided in the way of how the design process unfolded or the specific tensions that emerged between computational and interpretivist approaches.

Thus, this paper fills a gap by connecting these various threads. It describes how two interpretivist scholars, one in sociology and one in English literature, were incorporated into the design process. It offers examples of how each of these participants used the visualization tool in their own work. Finally, it articulates a series of tensions that emerged, partially as a result of emphasizing an interpretivist perspective, that can be used to help inform future design work.

3. DESIGN METHODS AND PROCESS

3.1. Training, Testing, and Case Studies: An Analogical Resemblance

The design process followed in this work bears an analogical resemblance to prior approaches in machine learning and in information visualization design. The process used here most closely resembles what has been termed visualization design studies (Sedlmair et al., 2012). Thus, this paper focuses less on presenting an entirely new design method and more on how an emphasis on supporting interpretivist scholarship reorients us toward many of the key concepts outlined in existing approaches.

3.1..1 Similarities

Machine learning (ML) often follows a paradigm referred to as train-test (Gelman & Hill, 2006). First, models are developed via iterative refinement on a training data set. Training here includes not only the fitting of the model to the data, but also decisions about which features to include in or exclude from the model. Second, the trained model is applied to a test data set to evaluate its performance. The model is never trained on the test data, though the results may lead to further development of the model on the training data. Performance on this test data set indicates how well the model works with data that were not used for training the model, for feature selection, or for fine tuning parameters. For these reasons, the training and testing data sets often come from different domains.

This standard ML approach bears some key similarities to visualization design and evaluation methods that make extensive use of case studies (Shneiderman & Plaisant, 2006) or design studies (Sedlmair et al., 2012). Case studies generally offer "detailed reporting about a small number of individuals working on their own problems, in their normal environment" (Shneiderman & Plaisant, 2006, p. 1). Just as training and testing data sets sometimes come from different domains, case studies are often conducted with individual experts across a variety of domains (there is, though, significant variance in how case studies are defined, conducted, and reported; see Isenberg et al., 2013). Similarly, just as an ML model is seen as superior when it performs well on multiple domains, a visualization design is preferable when it can be used by a variety of experts across different domains. A comparison can also be made with visualization design studies, a process in which "visualization researchers analyze a specific real-world problem faced by domain experts, design a visualization system that supports solving this problem, validate the design, and reflect about lessons learned" (Sedlmair et al., 2012, p. 2432)[1]. While design studies generally focus on a single domain, rather than looking across multiple domains, all these approaches also emphasize iterative refinement, with redesigns incorporating incremental feedback from, and observations of,

[1]The astute reader may notice a similarity between visualization design studies and research through design Zimmerman & Forlizzi (2014), but a full comparison of these approaches exceeds the scope of this paper.

a few usually expert users. Granted, key differences exist between train-test and case studies, such the methods and metrics for evaluating performance, case studies' emphasis on in-situ use, and others. Such differences notwithstanding, we cite this work to demonstrate how the design process here draws inspiration from established techniques both in machine learning and in visualization design.

For example, data scientists working in an ML training phase might attempt including or excluding specific features in a predictive model. Analogously, the work described here iteratively included or excluded different features and functionality from the visualization being developed. Similarly, our early designs were developed on a single corpus and then, in a step analogous to the ML testing phase, were applied to an entirely different corpus. Doing so helps determine how well the visualization performs, both on a corpus to which it was not applied while being developed and for a researcher who was not involved in the original design process.

3.1..2 Differences

That said, the emphasis on supporting interpretivist scholarship impacts many of the key notions traditionally applied in train-test (Gelman & Hill, 2006), in case studies (Shneiderman & Plaisant, 2006), and in visualization design studies (Sedlmair et al., 2012). For example, performance of the underlying ML model cannot be meaningfully assessed in isolation, using the kinds of quantitative metrics traditionally applied to natural language process (NLP) systems, such as F_1-score or area under the receiver operating characteristic curve (AUC). In ML, the move from the training to the testing phase often occurs once a model has achieved sufficient performance using such metrics. However, high performing models offer no guarantee of a useful visualization. Instead, performance must be assessed based on how well the visualization powered by the model supports expert users in the conduct of their work.

To some extent, such a strategy resembles that taken with case study (Shneiderman & Plaisant, 2006) and design study (Sedlmair et al., 2012) approaches. At the same time, the kinds of data analysis tasks in which interpretivist scholars are engaged differ from those in prior work using such approaches. Often, visualization case studies or design studies involve experts who aim to derive definitive insights about the nature of a data set or phenomenon (e.g., Kok et al., 2010; Seo & Shneiderman, 2006; Perer & Shneiderman, 2008) (see Isenberg et al., 2013; Sedlmair et al., 2012, for additional examples). In contrast, the scholars with whom we are designing seek to develop (potentially novel) interpretations of a text, supported by evidence from the text itself. Such interpretations differ from the kinds of objective insights prior visualization case studies aim to support. Throughout, this paper notes how such an emphasis on interpretation influenced both the visualization design and the underlying topic model.

Furthermore, the design and evaluation approaches described above (Shneiderman & Plaisant, 2006; Alexander et al., 2014; Klein et al., 2015; D'Ignazio & Klein, 2016) focus on situations with consistent underlying data. Although the visualization may be redesigned to highlight one or another aspect of that data, or the data may change dynamically over time, the nature or structure of the underlying data remain consistent. However, in the case of ML-based visualizations, the underlying model may be redesigned concomitantly with the visualization. Generally speaking, such model redesigns may include changes in feature selection or engineering, use of different classification algorithms, application of different loss/reward functions, testing alternative neural architectures, etc. Examples in the work presented here include curating an appropriate stop

word list, choosing the number of topics, generating suitable labels for each topic, and others described further below. Thus, our design process, in its consideration of the interplay between the visualization and the underlying model, necessarily differed from prior work focusing on either alone.

3.1..3 Desiderata

Design sessions with our expert collaborators were structured as think aloud protocols. At a high level, these collaborators were asked to provide feedback about which questions the current iteration of the visualization helped them address, as well as other questions that were raised but the visualization did not help address. Over the course of several sessions, our expert collaborators developed a list of research questions that they, as interpretivist scholars, would want the visualization to help them address.

- What are the *prevalent* topics? Put differently, what is this corpus about?
- How do the topics *relate* to each other?
- What other *relationships* within the corpus can be identified?
- How do the topics *change over time*?

The next subsection describes the two corpora used during this process.

3.2. Two Corpora

The visualization was initially designed during an iterative, multi-year project in collaboration between an HCI researcher (Author Baumer) and a sociologist (Author McGee). The corpus for this initial (a.k.a., "training") phase came from a collection of blogs written by parents with children on the autism spectrum (hereafter the ASD corpus). Selection of the blogs to include was guided by McGee, who is both an expert in this domain and a member of this blogging community. The corpus consists of 31,976 posts from 46 blogs, with a total of 8,136,071 (non-stopword) tokens. Posts span a twelve-year period from 2003 to 2015.

After several iterations on this ASD corpus, the rough shape of the visualization had been set, and co-design sessions focused more on minor details of the interaction. At this point, the visualization was tested by applying it to the diaries of Thomas Butler Gunn[2]. These diaries include 4077 separate entries (with 500,099 non-stopword tokens) written while Gunn, from England, lived and traveled in the US from 1849 through 1863. The early and middle portions focus on the Bohemians of New York City, while the latter years document the outbreak of the US Civil War. An English literature scholar (Edward Whitley) and a Senior Digital Scholarship Specialist (Rod Weidman), both familiar with this corpus, helped test the visualization design and generated some additional refinements, as noted below.

These two corpora bear some similarities. They both include observations of daily life events. They both chronicle the life of the writer and the lives of others. They both span a similar length of time. However, they also differ in many ways: multiple authors vs. a single author, modern vs. historical, blogs vs. hand-written diaries, 10,000s of documents vs. 1000s of documents. This

[2] `https://mohistory.org/collections/item/resource:103591` and `https://bit.ly/2XeDid5`

set of similarities and differences reduces the likelihood of overfitting the visualization to any one corpus.

4. RESULTS

This section presents separately the results that emerged from each of our two corpora, i.e., the ASD corpus and the Gunn corpus. In reality, there was some temporal overlap. As with the train-test machine learning paradigm on which this design process draws, application of the visualization to the "testing" data set, i.e., the Gunn corpus, revealed novel issues and opportunities. These revelations precipitated further iterative development with the "training" data set, i.e., the ASD corpus. For clarity, our work with each of the two corpora, and their impacts on the design, are presented separately.

4.1. Corpus I: ASD Parent Blogs

Our design process began with two existing topic modeling visualizations: jsLDA (Mimno, 2013) and the Dissertation Browser (Chuang, Ramage, et al., 2012). These were selected because source code was readily available[3] and/or McGee saw the visualizations as potentially useful for the ASD Parent Blogs. The resulting visualizations were simultaneously provocative yet unsatisfying. They demonstrated the potential utility that topic model visualizations could provide for interpretivist research, but they were unable to answer many of the questions they raised.

Prompted by these initial prototypes, we collaboratively developed a set of desiderata for a topic model visualization. These desiderata are articulated in the form of research questions that an interpretivist scholar would likely wish to address using topic modeling, as described above. The ultimate design of Topicalizer is specifically tailored to aid in addressing these questions. Thus, each subsection of the results describes how the visualization design addressed, or in some cases did not fully address, each of these questions. Throughout this description, we both attend to the ways that an emphasis on supporting interpretivist scholarship guided various design decisions and contrast these design decisions with prior topic model visualizations.

4.1..1 What are the *prevalent* topics?

At a very basic level, McGee wanted to know what these bloggers talked about (cf. "overview first" from Shneiderman (1996)). To that end, Topicalizer starts with a corpus view. This view presents the topics as a Venn diagram (rendered using Venn.js (Frederickson, 2019)). The size of each circle represents the number of documents in which each topic occurs, i.e., the number of documents where the topic proportion is greater than 0.05 (i.e., 5%) for that topic. This size provides a rough estimate for how much of the corpus is about each topic. The size of each intersection indicates the number of documents in which each pair of topics occurs (see Figure 1).

Two adjustments were made to the base Venn.js layout. First, to ease layout computation, only pair-wise relationships are considered. Our initial explorations suggested the relationships between sets of two topics as more generative and more readily interpretable then sets of three, four, or more topics. Second, the size of each intersection is scaled down linearly (here, divided by 100).

[3]Two other visualizations, TOME (Klein et al., 2015) and Serendip (Alexander et al., 2014), also seemed attractive, but at the time of commencing the work presented here their source code was not publicly available.

Initially, there was so much co-occurrence between topics that Venn.js layed out all circles virtually on top of one another. Linearly scaling down the sizes of the intersections still gives an approximation of how often each pair of topics co-occurs while simultaneously increasing legibility.

[- Figure 1 about here. -]

This final view was arrived at after several iterations. For example, an earlier version used a chord diagram (Figure 2). The size of each chord around the circle indicated the number of documents about each topic. The size of the ribbons between the chords indicated the number of documents in which each pair of topics occurred. Hovering over a single chord provided a pop-up with the highest probability words from the topic.

Initially, this chord diagram was very attractive as a high-level corpus overview. Perhaps the most interesting aspect in the initial prototypes mentioned above, based on jsLDA (Mimno, 2013) and on the Dissertation Browser (Chuang, Ramage, et al., 2012), was the ability to see relationships among topics. The chord diagram put those relationships front and center. However, as McGee put it during a design session, the primary interface needed to be "less daunting in terms of legibility." That is, it needed to provide immediate scaffolding to offer a quick, at-a-glance overview of the prevalent topics in the corpus.

[- Figure 2 about here. -]

The Venn diagram version of the corpus view accomplished this goal much more readily. With the chord diagram, the largest topics (i.e., those occurring in the most documents) drew the most immediate attention. Often, though, these topics turned out to be lower coherence (Lau et al., 2014) and less helpful in terms of helping scaffold interpretations. For example, with the chord diagram, the topic "not, time, get"[4] drew significant attention due to its size, but it did not reveal particularly interesting insights. Placing the words for each topic directly on the Venn diagram allowed a viewer's eyes, rather their pointing device, to identify quickly topics of potential interest. For example, with the Venn diagram, McGee never mentioned the larger topics, instead moving quickly to topics of conceptual interest, such as "police, story, mother" or "autism, study, children." Similar results occurred with the Gunn corpus, described further below.

This non-linearity is also an important component of interpretivist analysis (Charmaz, 2006; Lofland et al., 1971 (2005); Glaser & Strauss, 1967). The ability to follow conceptual threads of interest, rather than a systematic, pre-defined ordering, is crucial. Supporting such non-linearity motivated the use of the Venn diagram and of the chord diagram as the primary corpus overview, since both forms hinge on circles and provide a less strict ordering. This contrasts with prior systems, such as Serendip (Alexander et al., 2014) or Termite (Chuang, Manning, & Heer, 2012), which use more rectilinear layouts. In some ways, the Venn diagram has a non-linearity similar to that of TOME (Klein et al., 2015), though for the entire corpus rather than only for a subset of topics.

Some topics, such as "vaccine, vaccines, children", are readily comprehensible. Interpreting other topics, such as "may, point, fact," requires skimming documents with the greatest proportion of that topic. To that end, clicking on a circle in the Venn diagram shows the documents with the highest proportion of that topic (Figure 3, left). Clicking the title of a document shows

[4]This paper uses "quotes" for the names of topics, derived from the topic's high probability words, and SMALL CAPS for researchers' manually-assigned topic labels.

a breakdown of all the topics in that document (Figure 3, right), while clicking the URL opens the original post a new tab. Many other topic visualizations provide similar functionalities (e.g., Chuang, Ramage, et al., 2012; Mimno, 2013; Chaney & Blei, 2012).

[- Figure 3 about here. -]

However, the ability to look at the high-proportion documents for a topic takes on a different significance in the context of interpretivist work. Computational techniques such as topic modeling are sometimes described as "distant reading" (Rhody, 2017; Clement et al., 2013), in contrast to the "close reading" of texts typical in humanistic and qualitative sociological inquiry. Juxtaposing the high level description of a topic with excerpts where the topic occurs can help to span this division between distant vs. close reading.

For example, in the case of "may, point, fact," many posts deal with what is known (or not known) about autism and how. One such post discusses a finding in experimental particle physics where a neutrino was observed traveling faster than the speed of light. The blogger points out that no one immediately decried the result as something we know cannot be true, nor did anyone attack the personal credibility of the scientists involved, nor did anyone claim that the single result offers conclusive proof of anything. Instead, the response suggested that the observation needed to be approached with the same scientific rationality as all empirical observations are. The blogger then contrasts this with the responses that research on autism and vaccines garners, which, according to the blogger, do often involve hyperbole and personal attacks.

Many other posts with high proportions of this topic deal with similar questions about the processes of knowledge production. Thus, this was dubbed the EPISTEMIC topic. McGee had suspected that both epistemic and ontological issues permeated this corpus. The both the status of autism – is it a neurological pathology that warrants treatment and accommodations? a side effect from poor government regulation of pharmaceutical companies? a natural genetic variation to be celebrated? etc. – and what counts as valid means of knowing about this status are frequent points of complex tensions among ASD communities. Seeing this topic both helped to confirm that suspicion and helped direct McGee toward relevant documents.

In the case of some topics, the high probability words seemed interpretable, but certain subtleties were not captured by the label. For example, many documents with a high proportion of "vaccine, vaccines, children" specifically argue that early childhood vaccination causes autism. Examples range from criticism of routine vaccination schedules, to listing vaccines that contain thimerosal, which is the mercury-based preservative thought to give vaccines their autism-inducing properties. This ANTI-VAX standpoint was visible only after reviewing the documents in this single topic view.

Thus, providing excerpts of the documents, along with links to the original sources, made it easier to interpret what each topic was about. Furthermore, a viewer can examine a given topic in whatever depth is necessary or relevant for them to determine, as it were, what these documents are about.

4.1..2 How do the topics *relate* to each other?

Early explorations with jsLDA (Mimno, 2013) revealed relationality as a key interest. Being able to know, for example, that a post included both the EPISTEMIC and ANTI-VAX topics provides richer opportunities for interpretation than knowing that a post included either single topic. However,

observation of two topics co-occurring in one document quickly gave rise to questions about the overall relationships between different topics.

Several views in Topicalizer reveal such relationships among topics. As noted above, the single document view (Figure 3, right) shows which topics co-occur in a specific document. Furthermore, the corpus view (Figure 1) shows how often pairs of topics co-occur by the size of the intersections. Clicking on one of those intersections enables a more detailed comparison.

First, a scatterplot shows all posts in which both topics occur. The horizontal position of each document (i.e., point), shows the ratio of the proportions for the two selected topics. Documents near the middle have equal proportions of each topic, while documents toward the left or right side have a greater proportion of one topic or the other. The vertical position shows the sum of the proportions for the two topics selected, i.e., how much the document is about both topics.

[- Figure 4 about here. -]

The shape of this scatterplot can depict an overall relationship of the two topics. For example, Figure 4 shows that many ANTI-VAX documents contain small proportions of the EPISTEMIC topic. This relationship is shown by the density of dots along the right side of the scatterplot. Conversely, documents that are primarily about the EPISTEMIC topic less often include small proportions of the ANTI-VAX topic, as shown by the sparsity of points on the left side of the scatterplot. Put succinctly, this view shows that vaccines often raise epistemic issues, but not all epistemic issues deal with vaccines. The ability to comprehend this relationship at a glance contrasts with the dust-and-magnets approach used in TopicViz (Eisenstein et al., 2012), which requires sustained user interaction to get a sense for relationships among topics.

4.1..3 What *relationships* within the corpus can be identified?

The ASD corpus includes several dozen different blogs. One interesting question deals with how these blogs resemble or differ from one another. The topics present in a single post may or may not be as helpful as understanding an aggregate of a blog as a whole.

Thus, Topicalizer presents the ability to aggregate documents, for example, showing average topic proportions across all posts for a given blog. This view was designed to resemble the single document view (Figure 3, right), with a similar ability to transition to individual topics. This view can help reveal an individual blog's overall topics, even if some of those topics rarely occur together in the same post.

Topicalizer also offers the ability to compare multiple blogs. To do so, a force-directed layout is generated based on cosine similarity of the average topic proportions for each blog. Each blog is represented as a miniature version of the donut seen in the single aggregate view. Hovering over a blog's donut shows the blog's name in a pop-up, as well as thickens the network edges to that blog's immediate neighbors (Figure 5). This view is accompanied by a legend, similar to Figure 3 (right), which is omitted here for space. Any node in the network can be clicked to reach the single aggregate view for that blog.

[- Figure 5 about here. -]

This view immediately shows relationships among different blogs in terms of the topics they discuss. It also maintains the non-linearity and circular-based layouts used in other views. Within

this network (Figure 5), McGee was able to identify several meaningful subgroups. For example, the outliers near the right were neurodiversity advocates. The visualization grouped these blogs together because of their similar proportions of the EPISTEMIC topic and of the topic "autism, study, children." This later topic includes significant discussion of scientific studies on the pathogenesis and prevelance of autism. Many posts interpret these studies to suggest that autism is less of a mental illness and more a natural genetic variation – an integral part of the neurodiversity stance on autism. Put differently, these blogs were not necessarily linked because of a single neurodiversity topic, but because of multiple topics that occurred in similar proportions across these blogs. This instance offers just one example of how this aggregate view can reveal relationships within the corpus.

4.1..4 How do the topics *change over time?*

Many prior topic modeling visualizations include components designed specifically to represent (change over) time (e.g., Chuang, Ramage, et al., 2012; Eisenstein et al., 2012; Mimno, 2013; Klein et al., 2015). Initially, we developed an amplification of the time series plots from jsLDA (Mimno, 2013). These views showed the raw counts of documents with varying proportions of a given topic. Hovering shows a popup with numerical counts for the numbers of documents at each threshold (Figure 6).

[- Figure 6 about here. -]

Ultimately, though, this view was removed. In the few cases where it proved interesting, it rarely added additional insight. For example, the HOLIDAYS topic shows a clear periodicity, spiking every December (Figure 6, top). Similarly, there is a topic about Andrew Wakefield, whose article in the Lancet that linked early childhood vaccination with autism was later retracted Wakefield et al. (1998). This WAKEFIELD topic has a clear spike immediately after the retraction of the original paper, as well as another spike one year later. Such cases largely confirmed what was already known, rather than providing novel insights or scaffolding additional interpretation.

Furthermore, not every corpus has a meaningful temporal dimension. Topicalizer is intended to be a more general purpose visualization to support interpretivist analysis, and not all corpora include temporal metadata. Thus, it omits a specifically time-based view. Instead, time is treated as a different kind of aggregate. Just as documents can be grouped based on the blog from which they come, documents can also be grouped based on the month or year in which they were written. Thus, the aggregate views above can similarly be used to examine changes over time but in a non-linear fashion that aligns with the remainder of the visualization's design sensibility. The application was developed more fully after applying the visualization to the Gunn corpus, described further below.

4.1..5 Summary and Usage

The version of Topicalizer emerging from this initial work with the ASD corpus provides a variety of different views and interactions. These are summarized in Figure 7, which shows how each of the views connects to other views.

[- Figure 7 about here. -]

This flow among views arose from the confluence of three different factors. First, it follows the "visual information seeking mantra" (Shneiderman, 1996). It first provides an overview of all topics. It then allows a user to zoom in on particular topics or documents, filtering out others. Finally, it provides additional details, in many cases using pop-ups.

Second, it was constrained by the elements that are visible on the screen at any given time. For example, only the corpus view shows relationships among different topics. Thus, the scatterplot depicting the relationship between any pair of topics (Figure 7, center) can only be reached from the corpus view (Figure 7, top). In contrast, both individual topics and individual documents (or aggregates of documents) are visible in several different views. Thus the single topic view (Figure 7, left) and the single document (or aggregate) view (Figure 7, bottom) can be reached from multiple different views.

Third, and perhaps most importantly, the flow among different views co-evolved with the way that McGee used the visualization. Our initial prototypes, based on jsLDA (Mimno, 2013) and the Dissertation Browser (Chuang, Ramage, et al., 2012) were designed with a logically ordered, mostly linear flow. For example, our adaptation of jsLDA focused on the user selecting a topic, reading representative documents to determine what the topic is about, finding other topics with which it co-occurs, and examining how it ebbs and flows over time. This linear ordering provided a logical flow to the interface.

However, this logical ordering in early prototypes also caused issues. First, each of the views needed to be explicitly selected. McGee was required to remember and internalize how each of these views relates to one another, as well as when to switch to each. Second, because the flow was mostly linear, is had both a start and an end point. As a result, McGee often felt stuck in a "dead end," needing to repeat the process with each new topic.

In contrast, Topicalizer provided a much less linear prescription of how it should be used. The first two views to be implemented were the single topic and single document views (Figure 3). In an early co-design session, McGee would use these two views in a curiosity-driven wandering. Viewing a given topic would reveal a list of documents. When one of these documents seemed particularly interesting, the other topics in that document were then inspected, and so on. This oscillation between topics and documents would generate higher level questions about the relationships between different topics, and about the relationships among different blogs. It was for this reason that every view offers the ability to return to the initial corpus view. Thus, the flow among the different views offers another example of how the visualization during this initial design phase to come more in line with interpretivist analytic approaches.

Furthermore, we have noted above several places where the circular aesthetic of Topicalizer differs from the more rectilinear aesthetic of many prior topic modeling visualizations (Chuang, Ramage, et al., 2012; Alexander et al., 2014; Chaney & Blei, 2012; Mimno, 2013). This subsection illustrates how that non-linear approach emerged from, and then was consciously incorporated into, the design. The oscillation described above, between higher level themes (topics) and instances of those themes (documents), bears a strong resemblance to the continuous comparison of individual instances in the data with broader themes, a hallmark of the qualitative, interpretivist research methods that Topicalizer is designed to support (Charmaz, 2006; Glaser & Strauss, 1967; Lofland et al., 1971 (2005)).

4.2. Corpus II: The Gunn Diaries

As is common practice when testing a visualization across multiple data sets (Shneiderman & Plaisant, 2006), the Gunn Diaries both resemble and differ from the ASD Parent Blogs in various ways. As described above, the corpus is about an order of magnitude smaller (4077 diary entries, vs. 31,976 blog posts) and is all written by a single author. In addition, Whitley and Weidman were more familiar with this corpus than McGee was with the ASD corpus. As a result, their use of Topicalizer's different views generally oscillated between two modes: using the visualization to confirm their intuitions, and surprise at patterns and trends they had not anticipated. For consistency and comparison, these results are organized under subheadings that align with those in the preceding section.

4.2..1 What are the *prevalent* topics?

Preprocessing

As mentioned above, the design of a topic modeling visualization includes more than the interactive graphical components. It also includes details of how the model is trained on the data. Several of those details were refined when applying Topicalizer to the Gunn corpus.

First, for the ASD corpus, many of the topics from early prototypes consisted primarily of proper names. These were often names of, or pseudonyms for, family members of a specific blogger. Although some work suggests that such stopwords can be excluded after inferring topics (Schofield et al., 2017), we found that leaving the stopwords in the corpus interfered too much with the inferred topics[5]. Since a major goal was understanding thematic patterns across blogs, such names were added to the stopword list for that corpus. Similar prior work has at times developed fairly extensive custom stopword lists (Underwood & Goldstone, 2013; Jockers, n.d.).

The Gunn corpus, however, raised no such issue. Rather than being associated with a single author, as in the ASD corpus, such topics in the Gunn corpus revealed groups of people who inhabited similar social circles and whose activities together were documented by Gunn. Due to the interpretive possibilities such topics enabled, and with Whitley's advice, we left such names off the stopword list.

Second, a topic model must be fit using a predetermined number of topics. When working with the ASD corpus, we had manually tested different numbers of topics, finding that 50 topics provided a sufficient degree of granularity to be meaningful without having redundant topics. For the Gunn corpus, we similarly fit the topic model with different numbers of topics. Rather than manually inspecting them, we instead used techniques similar to Baumer et al. (2015) to cluster across these multiple sets of topics. Whitley determined that having 70 topics allowed for all the interesting themes to emerge without having multiple, redundant topics.

Later on in the design process, we experimented with using coherence to determine the number of topics. This numerical calculation, based on normalized pointwise mutual information, has been shown to correlate strongly with human judgments of topic quality (Lau et al., 2014; Lau & Baldwin, 2016). Thus, one can select the number of topics that maximizes average coherence. After testing several different numbers of topics for both corpora in this fashion, we found that the manually selected numbers of topics had the highest or near-highest coherence values. Thus,

[5] At the time this work was conducted, techniques for biasing topic models away from such known structure (Thompson & Mimno, 2018) had not yet been published

while the process for choosing a number of topics evolved, the results of that process remained roughly the same.

Third, both the ASD corpus and Gunn corpus included some documents that were less relevant to the research questions of interest. As mentioned above, McGee iterated repeatedly on the list of which blogs to include in the corpus. Some were excluded because, e.g., they pertained more to mental health in general than specifically to autism (parenting). Similarly, Gunn's diaries include a wide variety of additional material – newspaper clippings, loose leaf notes, wood block prints, poetry, etc. – inserted between the pages of the physical diary volumes. After much discussion, Whitley and Weidman decided that they were interested primarily in Gunn's observations in his own words. Thus, such supplemental content was excluded from the topic modeling analysis.

Such preprocessing steps tend to be highly corpus-specific. For example, optical character recognition (OCR) will sometimes identify the "long s" from older English books as an "f" (Springmann et al., 2014). Thus, we suspect that while some elements of this preprocessing step are fairly generic, others will need to be adapted to the unique demands for any data set.

Visualization

In their prior analyses, Whitley and Weidman had focused primarily on the middle years of Gunn's diaries. This choice was driven by Whitley's research interest in the American Bohemians of antebellum New York City. They were aware that Gunn had spent time covering the outbreak of the American civil war, but they had fewer expectations about the topics that should emerge from those portions of the corpus.

The choice to leave in proper names proved informative. For example, Gunn kept accurate records of the people who were present at the various social gatherings he attended. One topic included "cahill, boweryem, shepherd, miss, boley, phillips, maguire," and other names of Bohemian society who frequently held salon meetings. Such topics, in part, served to confirm the manual analyses Whitley and Weidman had previously done, and to confirm that the topic model identified patterns that were analytically informative.

However, some individual names occurred in multiple topics. For example, a second topic included "cahill, gun, bob, morris, ledger, arnold, drunk, shepherd." In both cases, "cahill" referred to Frank Cahill, renowned both for his presence in salon society and, as Whitley put it, his "drunken escapades." Indeed, this second topic pertained exactly to that latter theme. An excerpt from the second ranked document for this topic reads:

> "Corbin, Gun and Cahill came up in the evening, the two latter decidedly drunk, in spite of which, at Corbin's solicitation, we must needs go out to Mac Pyke's for more liquor, two thirds of which I contrived to spill on the floor, easy enough to do when your companions are inebriated. A dreary hour thus, then between two staggerers, home, Cahill zig-zagging his way across the street to his boarding-house." `https://bit.ly/31zj4Ko`

Thus, in this corpus, topics with a large number of proper names capture a combination of *which people* associate with one another and *in what contexts* they do so. This passage also demonstrates Gunn's disdain for many of the individuals he discusses, a point revisited below.

Another theme of interest dealt with Gunn's mental health. During an initial testing session, Weidman commented that their reading of the Diaries had given the impression that Gunn might

have suffered from depression. At this point, Weidman scanned through the list of topics and quickly identified one that seemed related to mental health: "sick, head, ill, nervous, weary, miserable, mind, horrible, lonely, doctor." While some of the representative documents dealt with ailments of the body, many focused on ailments of the mind. For example:

> "The whole day was like a relapse into my old nervous horror, and I, utterly lonely and miserable. Getting to bed, I lay awake for a long time, in pain of body and despondency of mind." `https://bit.ly/2XcJ0wi`

Seeing this apparent DEPRESSION topic, Whitley confirmed that Gunn seemed to be obsessed with William North, another writer who had committed suicide. Whitley then wanted to explore the literary implications further. Perhaps depression was common among the Bohemians? Perhaps a literary connection could be traced from Edgar Allen Poe's melancholic aesthetic to the Bohemians of antebellum New York? This example demonstrates how Topicalizer helped both to confirm and to further explore the researcher's initial impressions about potential themes and patterns.

4.2..2 How do the topics *relate* to each other?

Upon discovering the above-mentioned depression topic, both Weidman and Whitley were interested to see the other topics that co-occurred with it. One such topic was difficult to interpret based on the high probability words: "told, didn't, fellow, own, course, usual, story, heard." Upon inspection, this topic focused primarily on Gunn's judgment of others' character. For example, from the document with the third largest proportion of this topic:

> "Will Waud is just simply a selfish, conceited loafer. I suppose he married the girl on his usual principle of action – that of denying himself no indulgene he could come at. He'll neglect, perhaps ill-use, or desert her, if he grow tired of his plaything." `https://bit.ly/2KSjYf1`

Most entries for this topic contain similar negative statements resembling GOSSIP. The fact that GOSSIP co-occurred with DEPRESSION was not seen as particularly informative.

However, Whitley and Weidman proceeded to identify the other topics that co-occurred with this GOSSIP topic. They found many connections with the topics described above based on specific social groups. Incidentally, they noticed that GOSSIP co-occurs more often with the topic about Cahill's "drunken escapades" than with the topic about Bohemian salon society.

During these design sessions, Whitley and Weidman rarely investigated the question of *how* two topics co-occurred. That is, they did not spend time reviewing individual documents containing both topics. Instead, they were more interested in *whether* different topics co-occurred. Similar to McGee, these co-occurrences were used to facilitate interpretation of what each topic meant, as with the preceding GOSSIP example.

For this reason, Whitley and Weidman somewhat lamented the transition from the chord diagram (Figure 2) to the Venn diagram (Figure 1) for the corpus view. While the Venn diagram offered a parsimonious depiction, only showing the most prevalent pairwise topical co-occurrences, the ability to see every such occurrence in the chord diagram was also important for interpreting certain topics.

That said, the Venn view precipitated investigation of important topics that had previously gone unnoticed. For example, the topic "money, paid, pay, paying, board, debt, bill" quickly drew

Whitley's attention. While a few documents in this topic pertained to interpersonal debts, most dealt with the vast amounts of time, attention, and energy the Bohemians spent on getting paid. Whitley described narratives about of the "romance of Bohemianism," often crafted about these artists, that includes a discourse about nobly struggling to produce art against the demands of the marketplace. Such narratives, however, differ significantly from the actual "business of Bohemianism," the fact that members of these groups were always deeply invested in the question of where their next paycheck was coming from. At this point during the co-design session, Whitley stood up and took several photos of the visualization so he could later revisit the passages in Gunn's diary identified by the topic model and use them on a book project in which he was engaged.

4.2..3 What *relationships* within the corpus can be identified?

As mentioned, this corpus was written by a single author, which precludes identifying relationships among documents by different authors. Instead, Whitley and Weidman focused on comparisons among the different years of Gunn's diaries.

4.2..4 How do the topics *change over time*?

Both the historical relevance and the duration of Gunn's diaries (˜14 years) make temporality an important consideration. How do certain topics rise and fall over time? With what other historical or biographical events might they be associated? The approach to temporality again took the form of using the visualization either to confirm intuitions or to expose previously unidentified patterns.

Using the aggregate document view (Figure 8), Whitley and Weidman identified two significant shifts in Gunn's writing. The first, as expected, occurred in 1861, when he moved to the south to cover the outbreak of the American civil war. A distinct clique forms for the years 1855-1861. The year 1862 is connected to this clique only by its similarity to 1861.

[- Figure 8 about here. -]

However, this view revealed a second, unanticipated pattern. With the exception of 1850 and 1851, there was less topical similarity over the early years of Gunn's diaries. Whitley and Weidman had previously spent most of their time and efforts studying the period during the late 1850s, assuming that Gunn's early writing while in New York would be similar. Topicalizer suggested questioning that assumption, thus prompting further inquiry.

It is worth noting that these similarities are made more visible using the force-directed layout. If Topicalizer had represented change over time using time series plots, as in Figure 6, such clusterings of years would not be as readily apparent. As Whitley put it, this visualization design enabled them to investigate how various historical events, such as the civil war, affected American culture as seen through the lens of this diarist.

5. DISCUSSION

The results section above points out how the design of Topicalizer – both the final design and intermediate prototypes that were ultimately discarded – in some ways resembles and in other ways differs from prior work on topic modeling visualization (Eisenstein et al., 2012; Klein et

al., 2015; Goldstone, n.d.; Choo et al., 2013; H. Lee et al., 2012; Chuang, Manning, & Heer, 2012; Chuang, Ramage, et al., 2012; Chaney & Blei, 2012; Mimno, 2013). It also points out numerous ways that an emphasis on supporting interpretivist research methods (Charmaz, 2006; Corbin & Strauss, n.d.; Glaser & Strauss, 1967; Lofland et al., 1971 (2005)) influenced the design.

This discussion focuses on higher level issues and tensions that emerged across multiple phases throughout our design process. It emphasizes how working through these issues in this particular case – designing expressly for qualitative, interpretivist research methods – can provide insights and suggestions that may be more broadly applicable to systems at the nexus of AI/ML and HCI.

5.1. How Well Does It Work?

> "When a measure becomes a target, it ceases to be a good measure." – Goodhart's Law (Strathern, 1997, p. 308)

Machine learning techniques are often designed to perform according to quantitative metrics. Examples include precision, recall, accuracy, F-score, AUC, and others. Initially, topic modeling used perplexity on held out documents as a measure of how well the model fit the data. However, later work found almost no relationship between perplexity and human assessments of topic quality (Chang et al., 2009). A subsequent line of research has developed various coherence measures (Newman et al., 2010; Lau et al., 2014; Lau & Baldwin, 2016; Röder et al., 2015). Some form of a coherence measure is often used as a means for assessing overall quality of fit for a topic model (e.g., Schofield & Mimno, 2016).

However, as described above, we did not use such coherence measures when initially assessing our topic model (choosing stopwords, setting the number of topics, etc.), even though we later confirmed that the numbers of topics we had chosen yielded optimal or near-optimal coherence values. In some ways, a quantitative measure of model performance was not relevant here. To be sure, measures such as topic coherence can help, for example, in identifying instances where there maybe be a significant problem with the modeling. What is more important than the quantitative performance of the model, then, is how useful the model is for a human user.

5.1..1 Performance vs. Usefulness

Thus, one might wish to use different variants of the model to support human participants at performing a task. Various measures of human task performance could then be used to assess each variant's utility. However, quantitative measures of task performance often tell only a limited portion of the story when evaluating visualizations (Isenberg et al., 2013; Lam et al., 2012; Sedlmair et al., 2016). Indeed, with interpretivist work, the same kinds of performance metrics are not only unavailable, they would border on non-sensical.

Instead, more relevant here is the tool's ability to support the process of interpreting the meaning of a corpus of texts. Key to that process is the ability to formulate an evidenced argument. To what extent do various visualization designs (and different variants of underlying models) help these scholars craft evidenced arguments? This question could be investigated empirically: ask scholars to write arguments using different variants of the visualization, and then ask other scholars to evaluate those arguments. Doing so would not necessarily provide a purely quantitative comparisons. However, it could illuminate the qualitative differences in the kinds of evidence and arguments that one or another feature helps users to identify.

In many ways, this point mirrors arguments in HCI about the relationship between a system's usability and its usefulness (Greenberg & Buxton, 2008). While numerous techniques have been developed for usability evaluation, "usefulness is a very difficult thing to evaluate, especially by the usability methodologies common in CHI" (Greenberg & Buxton, 2008, p. 116). Within HCI, focusing primarily on usability can yield incremental changes but not revolutionary innovations. Similarly, within AI and ML, an abundant focus on performance assessment can lead to impressive improvements at specific tasks but cannot guarantee a model's utility to a human user. Indeed, significant recent work has highlighted issues that emerge in the sociotechnical gap (cf. Ackerman, 2000) between ML models that quantitatively perform well but produce confusing, misleading, biased, on nonsensical results (Eslami et al., 2015, 2016; DeVito, Hancock, et al., 2018; Bucher, 2017; Sweeney, 2013; Barocas & Selbst, 2016).

We suggest that future work should pursue blending these lines of thought. Incorporating the notion of usefulness into AI/ML evaluation can help move beyond quantitative performance metrics to more human-centered evaluation methods (e.g., Lertvittayakumjorn & Toni, 2019). While the work presented here leverages interpretivist scholarship as a context in which to demonstrate this point, such human-centered evaluation approaches likely apply in a broad variety of domains.

5.2. What Does This (Topic) Mean?

When first presented with the views in Topicalizer, all of McGee, Whitley, and Weidman began with selecting a given topic and trying to figure out what it was about. In some ways, this initial step resembles (intented or actual) usage patterns with other topic modeling visualizations (Chuang et al., 2015; Smith et al., 2018). However, it also differed, as topics fell broadly into one of four classes.

The first class of topics was relatively easy to understand. For example, it is fairly clear what topics such as "money, paid, pay" or "doctor, hospital, pain, medication" mean, simply by reading the high probability words.

The second class was less clear upon initial inspection. As described above, the EPISTEMIC topic was not initially comprehensible from its most frequent terms. However, the relevance of these terms became clearer after reading through several documents.

In the third class of topics, the top terms seemed to lack any coherence. This occurred in part because they were such common words, for example: "not, time, way, even, life, get, see." The most representative documents, however, suggest a common theme. One describes an autistic child graduating from high school. Another discusses the general changes that the arrival of a child precipitates in ones live. Another deals with beginning a new school year. Another still describes a planned change in occupation-related duties to enable a healthier work-life balance. In short, these documents all pertain to reflections on or around major life changes. However, even after seeing the documents and this theme, their connection with the topic's high probability words remains unclear.

A fourth class of topics adds further complexity. At first glance, both the ANTI-VAX topic and the Andrew WAKEFIELD topic deal with the connection between vaccines and autism. However, while navigating through Topicalizer, McGee noted that most of the ANTI-VAX documents argue for a causal relationship between vaccines and autism. The WAKEFIELD documents, on the other hand, argue against such a link and emphasize the retracted Wakefield et al. (1998) paper.

These findings carry implications for labeling in topic modeling. Topics can be labeled using

their most frequent words, as in this paper, or via other techniques, such as synthesizing a title from Wikipedia entries (Bhatia et al., 2016; Lau et al., 2011), or asking for labels from human annotators (Smith et al., 2017). Such techniques, though, are based on the notion that topics deal with *what* a document discusses, such as stock market trading or the olympic games (Lau et al., 2011). However, topics ANT-VAX or WAKEFIELD in the ASD corpus, as well as GOSSIP in the Gunn corpus, support the assertion that topic modeling identifies not topics but discourses (Rhody, 2013; Underwood, 2012, 2014). That is, topics may capture not only *what* documents are discussing but also *how* they are discussing it. Such cases suggest the opportunity to pursue developing alternative topic labeling techniques that emphasize these discursive aspects of topics.

5.2..1 Prescription vs. Interpretation

More broadly, this point highlights a tension between prescription and interpretation. That is, it demonstrates how algorithmic systems must strike a balance between explicitly stating what their results mean and intentionally leaving room for interpretation (Leahu, Schwenk, & Sengers, 2008; Leahu & Sengers, 2014; Leahu, Sengers, & Mateas, 2008). Indeed, prior work has identified complexities that can emerge with the social interpretation of algorithmic results (Gillespie, 2012; Eslami et al., 2016, 2015; Bucher, 2017; Ananny, 2011). For instance, the algorithm that Twitter uses to identify trending topics codifies a very specific, technical definition of what it means for a given topic to be popular. However, the high-order concept of "trending" can be interpreted in a variety of ways, some of which may differ drastically from the underlying algorithm (Gillespie, 2012). Similarly, social media news feed curation algorithms are likely optimized to increase clicks, dwell time, and other engagement metrics. However, users develop a broad diversity of strategies for navigating and interpreting what the content of these feeds reveals, both about themselves and about the nature of their relationships with others (Eslami et al., 2015, 2016; Bucher, 2017; DeVito, Birnholtz, et al., 2018).

To address this tension, the work presented here applies a strategy of linking higher level algorithmic patterns with specific instances thereof. In Topicalizer, as in other topic modeling visualizations (Mimno, 2013; Chaney & Blei, 2012; Alexander et al., 2014), the detailed view for a single topic shows how individual documents, and specific excerpts within documents, manifest the computationally identified topic. This approach can be seen as an echo of details-on-demand (Shneiderman, 1996).

In algorithmic systems, it becomes even more important to link computationally identified patterns with examples or instances thereof. Even in the ML-inspired participatory co-design approach taken here, collaborators in sociology and in English literature did not learn the mathematical details of topic modeling – nor should they be expected to. Put differently, we need not necessarily ensure that users and other stakeholders internalize a technically accurate mental model. Recent work on explainability in AI/ML may offer useful for techniques for exposing the inner workings of models (Samek et al., 2017; Došilović et al., 2018). At the same time, the goal of such explanations should not necessarily be ensuring that lay users or other stakeholders correctly understand the model's technical functioning. Instead, designers of algorithmic systems may be better served by balancing automated techniques that prescribe the meaning of computational results (e.g., automated topic labeling) with interactive representations that help scaffold interpretation of results (e.g., showing representative documents, or encouraging user-generated labels (cf. Smith et al., 2017)). Indeed, rather than *providing* an explanation or interpretation, there

is a valuable design space for techniques that *support* individuals and groups in forming their own interpretation, not only of how the system works but also of what it means. While a focus on intepretivist research methods highlights these issues, such approaches may prove useful beyond this specific application domain.

5.3. To Model or Not to Model?

These points, about what the model means or how it should be interpreted, draw attention to questions about what should be explicitly modeled, and what should perhaps not be. In the case of topic modeling visualization, this question has been explored both at the visualization level and at the model level.

First, most prior work on topic modeling visualization organizes results either in terms of documents or in terms of topics (Chaney & Blei, 2012; Mimno, 2013; H. Lee et al., 2012; Goldstone, n.d.; Dinakar et al., 2015; Chuang, Manning, & Heer, 2012; Chuang, Ramage, et al., 2012; Choo et al., 2013). Such designs should not be surprising, since topics and documents are two key entities in the mathematical formalisms of topic modeling (Blei et al., 2003).

However, the topical composition of a single document was often less interesting for McGee than that of each blog as a whole. Similarly, the topical composition for one specific diary entry from Gunn was less interesting for Whitley than overall patterns of co-occurrence or change over time. Similar to Topicalizer, TOME (Klein et al., 2015) visually supplements topic modeling results with metadata, e.g., the active editors of a given newspaper during different time periods. Although Klein et al. did not conduct "in situ studies of TOME as a component of real humanities research" (Klein et al., 2015, p. i140), the above results suggest that providing the ability to review patterns in regard to such metadata are key to supporting interpretability.

Second, one might consider employing models that account for and can explicitly represent the topic distributions of different authors (Steyvers et al., 2004; Rosen-Zvi et al., 2004; Roberts et al., 2014) (or academic departments (Chuang, Ramage, et al., 2012)). As noted above with TOME (Klein et al., 2015), capturing metadata – not only the author's name(s) but also gender, neurotypicality, political party affiliation, geographic location, etc. – affords significant analytic possibilities. However, explicitly modeling authors and/or other metadata raises issues in the corpora analyzed here, including multi-author blogs, guest posts, and extended quotes. We return to these issues below.

Stopword list curation demonstrates another issue at the modeling level. Following prior work (Underwood & Goldstone, 2013; Jockers, n.d.), the ASD corpus stopword list included many personal names found in the corpus. However, we also considered other alternatives. For example, one might replace a child's name with autistic_child, siblings' names with neurotypical_child, a partner's name with blogger's_partner, etc. Recent work also suggests more complex approaches, such as automatically identifying and subsampling words that are strongly associated with specific authors (Thompson & Mimno, 2018).

However, such substitutions would raise at least two issues. First, there is a potential slippery slope of substitutions, such as compassion_verb, vaccine_pejorative, etc. Inserting these substitutions into the data undermines the ability of topic modeling to identify patterns and relationships that a scholar using the tool might not expect or otherwise notice. Second, such substitutions would require significant commitment to analytic categories *a priori*. While different qualitative traditions differ in their embrace of *a priori* categories (see, e.g., Muller et al., 2016), such com-

mitments again undermine the ability of topic modeling to identify unexpected but informative patterns.

5.3..1 Explicit Modeling vs. Interpretive Flexibility

The above cases deal with questions of how exactly to represent data in the model. Such adjustments may allow better model fit in a mathematical sense. They may also offer additional resources for interpretation, such as a given blog's probability distribution over topics (i.e., the topics a given blogger is likely to write about). However, such adjustments can also introduce additional constraints. For example, every document would need to be assigned one or more authors, removing the possibility of anonymous documents (or requiring all anonymous documents to have been writted by a single author named Anonymous) and providing limited utility for single-author corpora.

As an alternative to explicitly representing such aspects in the model, the work presented in this paper leaned instead toward interpretive flexibility (Sengers & Gaver, 2006; Pinch & Bijker, 1987). Topicalizer, both the visualization and the underlying modeling, eschew explicit representations of many relevant concepts (vaccines, parenting, epistemology, etc.). For example, the system omits any means of automatic topic labeling (Bhatia et al., 2016; Mei et al., 2007; Lau et al., 2011). Instead, it exposes patterns through the presentation of carefully selected examples, where the selection methods emerged from the co-design process described above. This choice to avoid modeling or explicitly representing specific aspects of the data meant that user were afforded greater interpretive flexibility in terms of what the results mean.

Similar interpretive reorientations could be considered for other algorithmic systems. As a simple example, one could imagine turning a supervised learning problem into an unsupervised one. For example, consider algorithmic risk assessment (Angwin et al., 2016; Christin, 2017). A typical approach would treat each past case as recidivism or not, then train a classifier to predict future cases. Instead, one could use an unsupervised method to group the cases by underlying attributes, perhaps even those same attributes used as features in the supervised classifier (number of prior arrests, history of violent behavior, etc.). For a new case, the system would list past cases most similar to it, along with whether each of those cases resulted in recidivism. However, the system need not explicitly state whether the current case should be labeled high risk. Rather, it would leave the final decision to the user (likely, a judge). That human user could then take into account other factors, both of the case in question and of the comparison cases, to formulate their own risk assessment. In such a way, removing the explicit modeling of risk and recidivism from the system provides more room to interpret what the similarities between a given case and prior cases actually mean.

Such an emphasis on interpretive flexibility might not provide universal improvements. Emphasizing interpretation may slow down or prevent automated decision making, where high volumes of decisions must be made under limited time constraints, such as automated stock trading (`https://en.wikipedia.org/wiki/Automated_trading_system`) or resume screening (`https://hire.google.com/articles/resume-screening/`). On the one hand, it may be impractical or even impossible for human users to devote the time necessary to closely interpret every single data point. On the other hand, algorithmic tools that support but do not necessarily replace decision process may have significant room to be designed around interpretation. In any event, in situations where algorithmic systems work to support human activity, balancing explicit

modeling and interpretive flexibility may allow more room for questioning, and perhaps challenging, what the results of those systems mean.

6. CONCLUSION

Topicalizer, the system depicted in this paper, is not (necessarily) a final, finished product[6]. Similar to an ML model that achieves significant results in comparison to related techniques, Topicalizer offers significant differences, and in some cases improvements, beyond prior topic modeling visualization techniques. Furthermore, just as an ML model might be improved upon in subsequent work, we hope that future research will expand upon the unique affordances that Topicalizer provides for interpretivist research.

That said, this paper's primary contribution is an exploration of how an emphasis on interpretation plays out in the design and implementation of topic modeling visualization. The technical components of this work (visualization and topic modeling) draw on two subfields within computer science (human-computer interaction and natural language processing, respectively). Similarly, the design process presented here blends elements of participatory design with the training-and-testing paradigm of machine learning. Across each of these components, this paper highlights the shifts in the design, both process and product, that result from focusing on supporting interpretation. The discussion section then pulls out a number of high level tensions from this work that may be relevant to a variety of research and practitioners building other interactive algorithmic systems.

7. ACKNOWLEDGEMENTS

Thanks to the additional students who made various contributions to the coding for this visualization along the way: Yin Luo, Hannah Lambert, Alex Van Heest, and Rachel McCoog; to David Mimno for helpful conversations; and to the bloggers for sharing their stories. This material is based on work supported in part by the NSF under Grant No. IIS-1844901.

REFERENCES

Ackerman, M. S. (2000, September). The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility. *Human–Computer Interaction*, *15*(2-3), 179–203. doi: 10.1207/S15327051HCI1523_5

Alexander, E., Kohlmann, J., Valenza, R., Witmore, M., & Gleicher, M. (2014, October). Serendip: Topic model-driven visual exploration of text corpora. In *Conference on Visual Analytics Science and Technology (VAST)* (p. 173-182). Paris: IEEE. doi: 10.1109/VAST.2014.7042493

[6]Upon publication, the current version of the source code for Topicalizer will be made publicly available at `https://github.com/ericpsb/topicalizer`. This version will include data to generate and run the Gunn corpus visualization. However, it will exclude the ASD corpus due to concerns around the bloggers' identities and the sensitive nature of such data (cf. Baumer & McGee, 2019).

Ananny, M. (2011, April). The Curious Connection between Apps for Gay Men and Sex Offenders. *The Atlantic*.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May). Machine Bias. *Pro Publica*.

Bardzell, S. (2010). Feminist HCI: Taking Stock and Outlining an Agenda for Design. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)* (pp. 1301–1310). Atlanta, GA: ACM.

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*(3), 671–732.

Baumer, E. P. S. (2017). Toward Human-Centered Algorithm Design. *Big Data & Society*, *4*(2). doi: 10.1177/2053951717718854

Baumer, E. P. S., Guha, S., Quan, E., Mimno, D., & Gay, G. K. (2015). Missing Photos, Suffering Withdrawal, or Finding Freedom? How Experiences of Social Media Non-Use Influence the Likelihood of Reversion. *Social Media + Society*, *1*(2). doi: 10.1177/2056305115614851

Baumer, E. P. S., & McGee, M. (2019). Speaking on Behalf of: Representation, Delegation, and Authority in Computational Text Analysis. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)* (p. 8). Honolulu, HI: AAAI/ACM.

Baumer, E. P. S., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017, June). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology (JASIST)*, *68*(6), 1397-1410. doi: 10.1002/asi.23786

Berger, P. L., & Luckmann, T. (1966). *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Garden City, NY: Anchor Books.

Bhatia, S., Lau, J. H., & Baldwin, T. (2016, December). Automatic Labelling of Topics with Neural Embeddings. In *Proceedings of the International Conference on Computational Linguistics (COLING)* (pp. 953–963). Osaka, Japan.

Blei, D. M. (2012, April). Probabilistic Topic Models. *Commun. ACM*, *55*(4), 77–84. doi: 10.1145/2133806.2133826

Blei, D. M., & Lafferty, J. D. (2007a). A Correlated Topic Model of Science. *The Annals of Applied Statistics*, *1*(1), 17-35.

Blei, D. M., & Lafferty, J. D. (2007b). *Modeling Science.* http://www.cs.cmu.edu/~lemur/science/.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022.

Boehner, K., Vertesi, J., Sengers, P., & Dourish, P. (2007). How HCI interprets the probes. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)* (pp. 1077–1086). San Jose, CA: ACM. doi: 10.1145/1240624.1240789

Boyd-Graber, J., Hu, Y., & Mimno, D. (2017, July). Applications of Topic Models. *Foundations and Trends® in Information Retrieval*, *11*(2-3), 143-296. doi: 10.1561/1500000030

Boydstun, A. E., Card, D., Gross, J. H., Resnik, P., & Smith, N. A. (2014). Tracking the Development of Media Frames within and across Policy Issues. In *Annual Meeting of the American Political Science Association (APSA).* Washington, D.C..

Bucher, T. (2017). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, *20*(1), 30-44. doi: 10.1080/1369118X.2016.1154086

Burrell, J. (2016, January). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. doi: 10.1177/2053951715622512

Chaney, A. J. B. (2012). *Wikipedia Topics.* http://www.princeton.edu/~achaney/tmve/wiki100k/browse/topic-presence.html.

Chaney, A. J. B., & Blei, D. M. (2012). Visualizing Topic Models. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)* (p. 419-422). Dublin: AAAI.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems (NIPS)* (pp. 288–296). Curran Associates, Inc.

Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. London: SAGE Publications.

Chen, N.-C., Drouhard, M., Kocielnik, R., Suh, J., & Aragon, C. R. (2018, June). Using Machine Learning to Support Qualitative Coding in Social Science: Shifting the Focus to Ambiguity. *ACM Trans. Interact. Intell. Syst.*, *8*(2), 9:1–9:20. doi: 10.1145/3185515

Choo, J., Lee, C., Reddy, C. K., & Park, H. (2013, December). UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 1992-2001. doi: 10.1109/TVCG.2013.212

Christin, A. (2017, December). Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*, *4*(2), 2053951717718855. doi: 10.1177/2053951717718855

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization Techniques for Assessing Textual Topic Models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI)* (pp. 74–77). Capri Island, Italy: ACM. doi: 10.1145/2254556.2254572

Chuang, J., Ramage, D., Manning, C., & Heer, J. (2012). Interpretation and Trust: Designing model-driven visualizations for text analysis. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)* (pp. 443–452). Austin, TX: ACM.

Chuang, J., Roberts, M. E., Stewart, B. M., Weiss, R., Tingley, D., Grimmer, J., & Heer, J. (2015). TopicCheck: Interactive Alignment for Assessing Topic Model Stability. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)* (pp. 175–184). Denver, CO.

Clement, T., Tcheng, D., Auvil, L., Capitanu, B., & Barbosa, J. (2013, December). Distant Listening to Gertrude Stein's 'Melanctha': Using Similarity Analysis in a Discovery Paradigm to Analyze Prosody and Author Influence. *Literary and Linguistic Computing*, *28*(4), 582-602. doi: 10.1093/llc/fqt040

Concannon, S. J., Balaam, M., Simpson, E., & Comber, R. (2018). Applying Computational Analysis to Textual Data from the Wild: A Feminist Perspective. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)* (pp. 226:1–226:13). Montréal, QC: ACM. doi: 10.1145/3173574.3173800

Corbin, J., & Strauss, A. (n.d.). Grounded Theory Research: Procedures, Canons, and Evaluative Criteria. *Qualitative Sociology*, *13*(1), 3.

DeVito, M. A., Birnholtz, J., Hancock, J. T., French, M., & Liu, S. (2018). How People Form Folk Theories of Social Media Feeds and What It Means for How We Study Self-Presentation. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)* (pp. 120:1–120:12). Montréal, QC: ACM. doi: 10.1145/3173574.3173694

DeVito, M. A., Hancock, J. T., French, M., Birnholtz, J., Antin, J., Karahalios, K., … Shklovski, I. (2018). The Algorithm and the User: How Can HCI Use Lay Understandings of Algorithmic Systems? In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems (CHI EA)* (pp. panel04:1–panel04:6). Montréal, QC: ACM. doi: 10.1145/3170427.3186320

D'Ignazio, C., & Klein, L. F. (2016). Feminist Data Visualization. In *Workshop on Visualization for the Digital Humanities, at IEEE VIS Conference.* Baltimore, MD.

Dinakar, K., Chen, J., Lieberman, H., Picard, R., & Filbin, R. (2015). Mixed-Initiative Real-Time Topic Modeling & Visualization for Crisis Counseling. In *Proceedings of the ACM Conference on Intelligent User Interfaces (IUI)* (pp. 417–426). Atlanta, GA: ACM. doi: 10.1145/2678025.2701395

Došilović, F. K., Brčić, M., & Hlupić, N. (2018, May). Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (p. 0210-0215). doi: 10.23919/MIPRO.2018.8400040

Dove, G., Halskov, K., Forlizzi, J., & Zimmerman, J. (2017). UX Design Innovation: Challenges for Working with Machine Learning As a Design Material. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)* (pp. 278–288). Denver, CO: ACM. doi: 10.1145/3025453.3025739

Drouhard, M., Chen, N. C., Suh, J., Kocielnik, R., Peña-Araya, V., Cen, K., … Aragon, C. R. (2017, April). Aeonium: Visual analytics to support collaborative qualitative coding. In *IEEE Pacific Visualization Symposium (PacificVis)* (p. 220-229). Seoul: IEEE. doi: 10.1109/PACIFICVIS.2017.8031598

Ehn, P. (1988). *Work-Oriented Design of Computer Artifacts*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Eisenstein, J., Chau, D. H., Kittur, A., & Xing, E. (2012). TopicViz: Interactive Topic Exploration in Document Collections. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems (CHI EA)* (pp. 2177–2182). Austin, TX: ACM. doi: 10.1145/2212776.2223772

Eslami, M., Karahalios, K., Sandvig, C., Vaccaro, K., Rickman, A., Hamilton, K., & Kirlik, A. (2016). First I "Like" It, then I Hide It: Folk Theories of Social Feeds. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)* (pp. 2371–2382). San Jose, CA: ACM. doi: 10.1145/2858036.2858494

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., ... Sandvig, C. (2015). "I Always Assumed That I Wasn't Really That Close to [Her]": Reasoning About Invisible Algorithms in News Feeds. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)* (pp. 153–162). Seoul: ACM. doi: 10.1145/2702123.2702556

Foucault, M. (1972). *The Archaeology of Knowledge* (A. M. Sheridan Smith, Trans.). New York: Pantheon Books. doi: 10.1002/9780470776407.ch20

Frederickson, B. (2019, June). *Area proportional Venn and Euler diagrams in JavaScript: Benfred/venn.js.*

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gillespie, T. (2012, February). Can an Algorithm be Wrong? *Limn*, *1*(2).

Gillespie, T. (2013). The Relevance of Algorithms. In T. Gillespie, P. Bockzkowski, & K. Foot (Eds.), *Media Technologies.* Cambridge, MA: MIT Press.

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory*. Oxon and New York: Routledge.

Goldstone, A. (n.d.). *Topic Model Browser.* http://agoldst.github.io/dfr-browser/demo/.

Goldstone, A., & Underwood, T. (2014, November). The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us. *New Literary History*, *45*(3), 359-384. doi: 10.1353/nlh.2014.0025

Greenberg, S., & Buxton, B. (2008). Usability Evaluation Considered Harmful (Some of the Time). In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)* (pp. 111–120). Florence, Italy: ACM.

Isenberg, T., Isenberg, P., Chen, J., Sedlmair, M., & Möller, T. (2013, December). A Systematic Review on the Practice of Evaluating Visualization. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2818-2827. doi: 10.1109/TVCG.2013.126

Jockers, M. L. (n.d.). *Expanded Stopwords List.* http://www.matthewjockers.net/macroanalysisbook/expanded-stopwords-list/.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods & Literary History*. Chicago: University of Illinois Press.

Jockers, M. L., & Mimno, D. (2013, December). Significant themes in 19th-century literature. *Poetics*, *41*(6), 750-769. doi: 10.1016/j.poetic.2013.08.005

Klein, L. F., & Eisenstein, J. (2013, December). Reading Thomas Jefferson with TopicViz: Towards a Thematic Method for Exploring Large Cultural Archives. *Scholarly and Research Communication*, *4*(3).

Klein, L. F., Eisenstein, J., & Sun, I. (2015, December). Exploratory Thematic Analysis for Digitized Archival Collections. *Digital Scholarship in the Humanities*, *30*(suppl_1), i130-i141. doi: 10.1093/llc/fqv052

Knigge, L., & Cope, M. (2006, November). Grounded Visualization: Integrating the Analysis of Qualitative and Quantitative Data through Grounded Theory and Visualization. *Environment and Planning A: Economy and Space*, *38*(11), 2021-2037. doi: 10.1068/a37327

Kok, P., Baiker, M., Hendriks, E. A., Post, F. H., Dijkstra, J., Lowik, C. W. G. M., ... Botha, C. P. (2010, November). Articulated Planar Reformation for Change Visualization in Small Animal Imaging. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 1396-1404. doi: 10.1109/TVCG.2010.134

Kuhn, S., & Muller, M. J. (1993, June). Participatory Design. *Communications of the ACM*, *36*(4), 24–28. doi: 10.1145/153571.255960

Lam, H., Bertini, E., Isenberg, P., Plaisant, C., & Carpendale, S. (2012, September). Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Transactions on Visualization and Computer Graphics*, *18*(9), 1520-1536. doi: 10.1109/TVCG.2011.279

Lau, J. H., & Baldwin, T. (2016, June). The Sensitivity of Topic Coherence Evaluation to Topic Cardinality. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)* (pp. 483–487). San Diego, California: Association for Computational Linguistics.

Lau, J. H., Grieser, K., Newman, D., & Baldwin, T. (2011). Automatic Labelling of Topic Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 1536–1545). Portland, OR: Association for Computational Linguistics.

Lau, J. H., Newman, D., & Baldwin, T. (2014, April). Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 530–539). Gothenburg, Sweden: Association for Computational Linguistics.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., ... Alstyne, M. V. (2009, February). Computational Social Science. *Science*, *323*(5915), 721-723. doi: 10.1126/science.1167742

Leahu, L. (2016). Ontological Surprises: A Relational Perspective on Machine Learning. In *Proceedings of the ACM Conference on Designing Interactive Systems (DIS)* (pp. 182–186). Brisbane, Australia: ACM. doi: 10.1145/2901790.2901840

Leahu, L., Schwenk, S., & Sengers, P. (2008). Subjective Objectivity: Negotiating Emotional Meaning. In *Proceedings of the ACM Conference on Designing Interactive Systems (DIS)* (pp. 425–434). Cape Town, South Africa: ACM. doi: 10.1145/1394445.1394491

Leahu, L., & Sengers, P. (2014). Freaky: Performing Hybrid Human-machine Emotion. In *Proceedings of the ACM Conference on Designing Interactive Systems (DIS)* (pp. 607–616). Vancouver, BC: ACM. doi: 10.1145/2598510.2600879

Leahu, L., Sengers, P., & Mateas, M. (2008). Interactionist AI and the Promise of Ubicomp, or, How to Put Your Box in the World Without Putting the World in Your Box. In *Proceedings of the International Conference on Ubiquitous Computing (Ubicomp)* (pp. 134–143). Seoul: ACM. doi: 10.1145/1409635.1409654

Lee, H., Kihm, J., Choo, J., Stasko, J., & Park, H. (2012, June). iVisClustering: An Interactive Visual Document Clustering via Topic Modeling. *Computer Graphics Forum*, *31*(3pt3), 1155-1164. doi: 10.1111/j.1467-8659.2012.03108.x

Lee, T. Y., Smith, A., Seppi, K., Elmqvist, N., Boyd-Graber, J., & Findlater, L. (2017, September). The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*, *105*, 28-42. doi: 10.1016/j.ijhcs.2017.03.007

Lertvittayakumjorn, P., & Toni, F. (2019, November). Human-grounded Evaluations of Explanation Methods for Text Classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 5195–5205). Hong Kong, China: ACL. doi: 10.18653/v1/D19-1523

Lind, R. A., & Salo, C. (2002, January). The Framing of Feminists and Feminism in News and Public Affairs Programs in U.S. Electronic Media. *Journal of Communication*, *52*(1), 211-228. doi: 10.1111/j.1460-2466.2002.tb02540.x

Lofland, J., Snow, D. A., Anderson, L., & Lofland, L. H. (1971 (2005)). *Analyzing Social Settings*. Belmont, CA: Wadsworth.

Marathe, M., & Toyama, K. (2018). Semi-Automated Coding for Qualitative Research: A User-Centered Inquiry and Initial Prototypes. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)* (pp. 348:1–348:12). Montréal, QC: ACM. doi: 10.1145/3173574.3173922

Mei, Q., Shen, X., & Zhai, C. (2007). Automatic labeling of multinomial topic models. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 490–499). San Jose, CA: ACM.

Mimno, D. (2013). *jsLDA: In-browser topic modeling*.

Muller, M. J. (2012). Participatory Design: The Third Space in HCI. In A. Sears & J. A. Jacko (Eds.), *The Human-Computer Interaction Handbook* (Vol. 4235, p. 1061-1081). Hillsdale, NJ: Lawrence Erlbaum Associates.

Muller, M. J., Guha, S., Baumer, E. P. S., Mimno, D., & Shami, N. S. (2016). Machine Learning and Grounded Theory Method: Convergence, Divergence, and Combination. In *Proceedings of the ACM Conference on Supporting Group Work (GROUP).* Sanibel Island, FL.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic Evaluation of Topic Coherence. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)* (pp. 100–108). Stroudsburg, PA, USA: ACL.

Orlikowski, W. J., & Baroudi, J. J. (1991). Studying information technology in organizations: Research approaches and assumptions. *Information systems research*, *2*(1), 1–28.

Pang, B., & Lee, L. (2008, January). Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, *2*(1-2), 1–135. doi: 10.1561/1500000011

Passi, S., & Jackson, S. (2017). Data Vision: Learning to See Through Algorithmic Abstraction. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW)* (pp. 2436–2447). Portland, OR: ACM. doi: 10.1145/2998181.2998331

Perer, A., & Shneiderman, B. (2008). Integrating Statistics and Visualization: Case Studies of Gaining Clarity During Exploratory Data Analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 265–274). Florence, Italy: ACM. doi: 10.1145/1357054.1357101

Pinch, T. J., & Bijker, W. E. (1987). The Social Construction of Facts and Artifacts. In W. E. Bijker, T. P. Hughes, & T. J. Pinch (Eds.), *The Social Construction of Technological Systems* (pp. 17–50). Cambridge, MA: MIT Press.

Poursabzi-Sangdeh, F., Boyd-Graber, J., Findlater, L., & Seppi, K. (2016). ALTO: Active Learning with Topic Overviews for Speeding Label Induction and Document Labeling. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 1158–1169). Berlin: Association for Computational Linguistics.

Rhody, L. M. (2013, April). *Topic Modeling and Figurative Language.* http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/.

Rhody, L. M. (2017, May). Beyond Darwinian Distance: Situating Distant Reading in a Feminist *Ut Pictura Poesis* Tradition. *PMLA*, *132*(3), 659-667. doi: 10.1632/pmla.2017.132.3.659

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... Rand, D. G. (2014, October). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, *58*(4), 1064-1082. doi: 10.1111/ajps.12103

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. In *Proceedings of the ACM Conference on Web Search and Data Mining (WSDM)* (pp. 399–408). Shanghai, China: ACM. doi: 10.1145/2684822.2685324

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The Author-topic Model for Authors and Documents. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)* (pp. 487–494). Banff, AB, Canada: AUAI Press.

Samek, W., Wiegand, T., & Müller, K.-R. (2017, August). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models.

Schofield, A., Magnusson, M., & Mimno, D. (2017). Pulling Out the Stops: Rethinking Stopword Removal for Topic Models. In *Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL)* (Vol. 2, p. 432-436). Valencia, Spain: ACL.

Schofield, A., & Mimno, D. (2016, July). Comparing Apples to Apple: The Effects of Stemmers on Topic Models. *Transactions of the Association for Computational Linguistics*, *4*, 287-300.

Sedlmair, M., Isenberg, P., Isenberg, T., Mahyar, N., & Lam, H. (2016). Beyond Time And Errors: Novel Evaluation Methods For Visualization (BELIV 2016). In *Workshop at IEEE VIS Week.* Baltimore, MD.

Sedlmair, M., Meyer, M., & Munzner, T. (2012, December). Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2431–2440. doi: 10.1109/TVCG.2012.213

Sengers, P., Boehner, K., Mateas, M., & Gay, G. (2008, June). The Disenchantment of Affect. *Personal and Ubiquitous Computing*, *12*(5), 347-358. doi: 10.1007/s00779-007-0161-4

Sengers, P., & Gaver, B. (2006). Staying Open to Interpretation: Engaging Multiple Meanings in Design and Evaluation. In *Proceedings of the ACM Conference on Designing Interactive Systems (DIS)* (pp. 99–108). University Park, PA: ACM.

Seo, J., & Shneiderman, B. (2006, May). Knowledge discovery in high-dimensional data: Case studies and a user survey for the rank-by-feature framework. *IEEE Transactions on Visualization and Computer Graphics*, *12*(3), 311-322. doi: 10.1109/TVCG.2006.50

Shneiderman, B. (1996). The Eyes Have It: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages* (pp. 336–343). IEEE.

Shneiderman, B., & Plaisant, C. (2006). Strategies for Evaluating Information Visualization Tools: Multi-dimensional In-depth Long-term Case Studies. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization (BELIV)* (pp. 1–7). Venice, Italy: ACM. doi: 10.1145/1168149.1168158

Smith, A., Kumar, V., Boyd-Graber, J., Seppi, K., & Findlater, L. (2018). Closing the Loop: User-Centered Design and Evaluation of a Human-in-the-Loop Topic Modeling System. In *Proceedings of the ACM Conference on Intelligent User Interfaces (IUI)* (pp. 293–304). Tokyo: ACM. doi: 10.1145/3172944.3172965

Smith, A., Lee, T. Y., Poursabzi-Sangdeh, F., Boyd-Graber, J., Elmqvist, N., & Findlater, L. (2017, January). Evaluating Visual Representations for Topic Understanding and Their Effects on Manually Generated Labels. *Transactions of the Association for Computational Linguistics*, *5*, 1-15.

Springmann, U., Najock, D., Morgenroth, H., Schmid, H., Gotscharek, A., & Fink, F. (2014). OCR of Historical Printings of Latin Texts: Problems, Prospects, Progress. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH)* (pp. 71–75). Madrid, Spain: ACM. doi: 10.1145/2595188.2595205

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic Author-topic Models for Information Discovery. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 306–315). Seattle, WA: ACM. doi: 10.1145/1014052.1014087

Strathern, M. (1997, July). 'Improving ratings': Audit in the British University system. *European Review*, *5*(3), 305-321. doi: 10.1002/(SICI)1234-981X(199707)5:3⟨305::AID-EURO184⟩3.0.CO;2-4

Sweeney, L. (2013). Discrimination in online ad delivery. *Communications of the ACM*, *56*(5), 44–54. doi: 10.1145/2460276.2460278

Taylor, A. S. (2009). Machine Intelligence. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)* (p. 2109-2118). Boston, MA.

Thompson, L., & Mimno, D. (2018). Authorless Topic Models: Biasing Models Away from Known Structure. In *Proceedings of the International Conference on Computational Linguistics (COLING)* (pp. 3903–3914). Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Underwood, T. (2012, April). *What kinds of "topics" does topic modeling actually produce?*

Underwood, T. (2014). Theorizing research practices we forgot to theorize twenty years ago. *Representations*, *127*(1), 64–72.

Underwood, T. (2019). *Distant Horizons*. Chicago: University of Chicago Press.

Underwood, T., & Goldstone, A. (2013, September). List of stop words used in topic modeling journals, summer 2013.

Wakefield, A. J., Murch, S. H., Anthony, A., Linnell, J., Casson, D. M., Malik, M., … Walker-Smith, J. A. (1998, February). RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, *351*(9103), 637-641. doi: 10.1016/S0140-6736(97)11096-0

Yang, Q., Scuito, A., Zimmerman, J., Forlizzi, J., & Steinfeld, A. (2018). Investigating How Experienced UX Designers Effectively Work with Machine Learning. In *Proceedings of the ACM Conference on Designing Interactive Systems (DIS)* (pp. 585–596). Hong Kong: ACM. doi: 10.1145/3196709.3196730

Zade, H., Drouhard, M., Chinh, B., Gan, L., & Aragon, C. (2018). Conceptualizing Disagreement in Qualitative Coding. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)* (pp. 159:1–159:11). Montréal, QC: ACM. doi: 10.1145/3173574.3173733

Zimmerman, J., & Forlizzi, J. (2014). Research through Design in HCI. In J. S. Olson & W. A. Kellogg (Eds.), *Ways of Knowing in HCI* (pp. 167–189). New York: Springer.

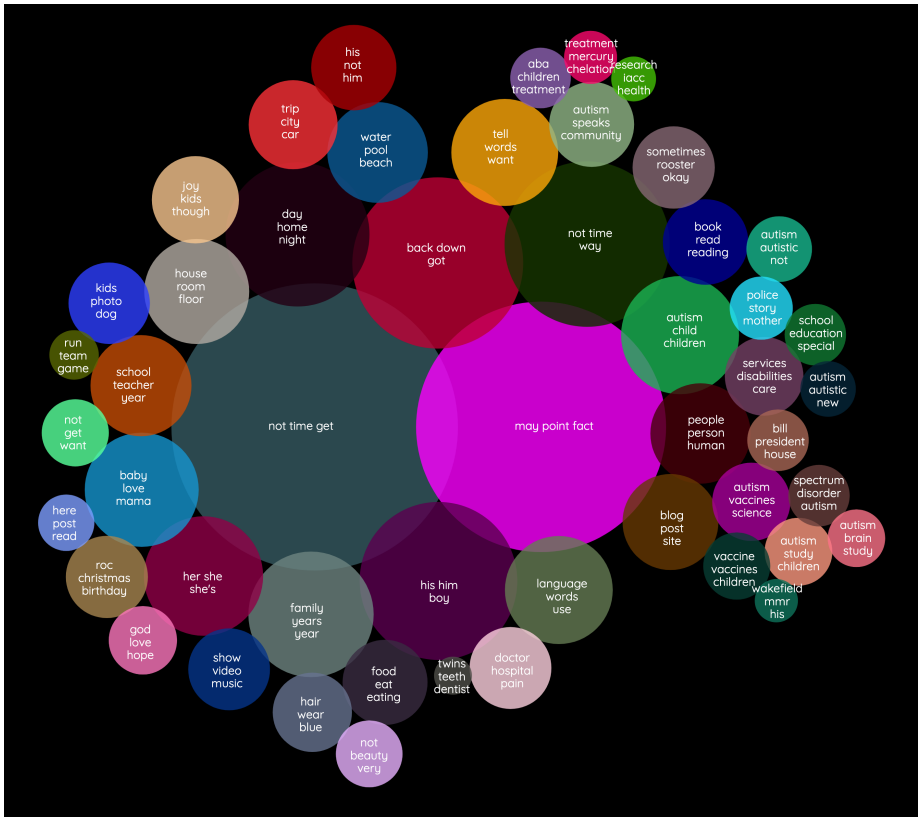

FIGURE 1. Topicalizer starts at the corpus view, a Venn diagram of topics. The size of each topic shows the number of documents in which that topic occurs, and the size of each intersection shows the number of documents in which each pair of documents occurs. Colors are randomly assigned from a palette generated to have as many visually distinct colors as possible. The corpus view is shown here using the ASD corpus.

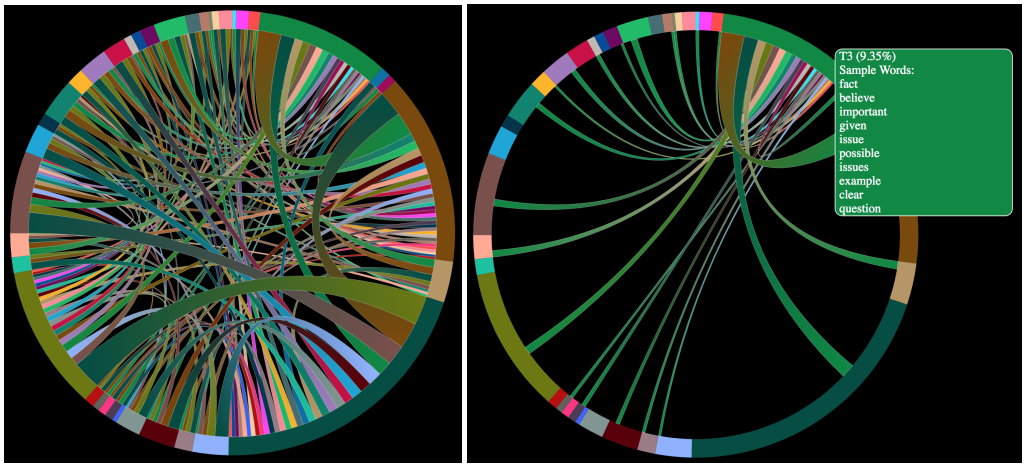

FIGURE 2. Early prototype corpus view, shown here using the ASD corpus. The size of each chord around the circle shows the number of documents in which that topic occurs, and the size of each ribbon shows the number of documents in which each pair of documents occurs. Hovering over a chord (above right) shows the words for that topic and hides all ribbons not connected to that topic.

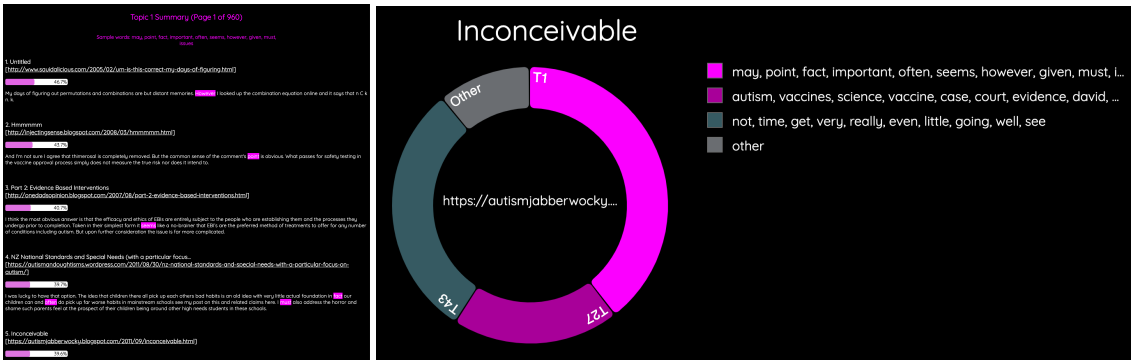

FIGURE 3. The single topic view (left) shows the documents with the highest proportion of a given topic. Clicking the document's title will open the single document view (right), which shows the title and the topic proportions for all the topics in a document. Here, clicking a topic will go to the single topic view. Topics that occur in proportions less than 0.05 are grouped into "other." In both views, clicking a URL will open a new tab with the original blog post.

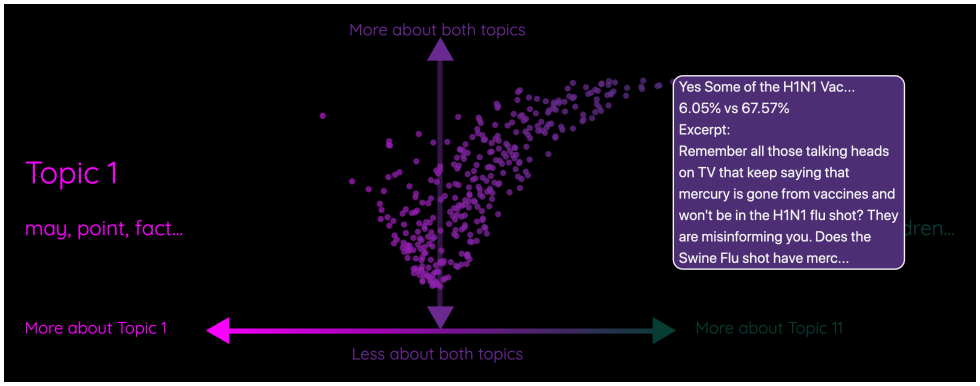

FIGURE 4. Topic comparison view, showing all documents that include both the EPISTEMIC topic (T1, left side) and the ANTI-VAX topic (T11, right side). The position of each dot indicates the each document's total proportion of the two topics (top to bottom) and the relative proportion of the two topics (left to right). Here, the user has hovered over the point for the document most about T11.

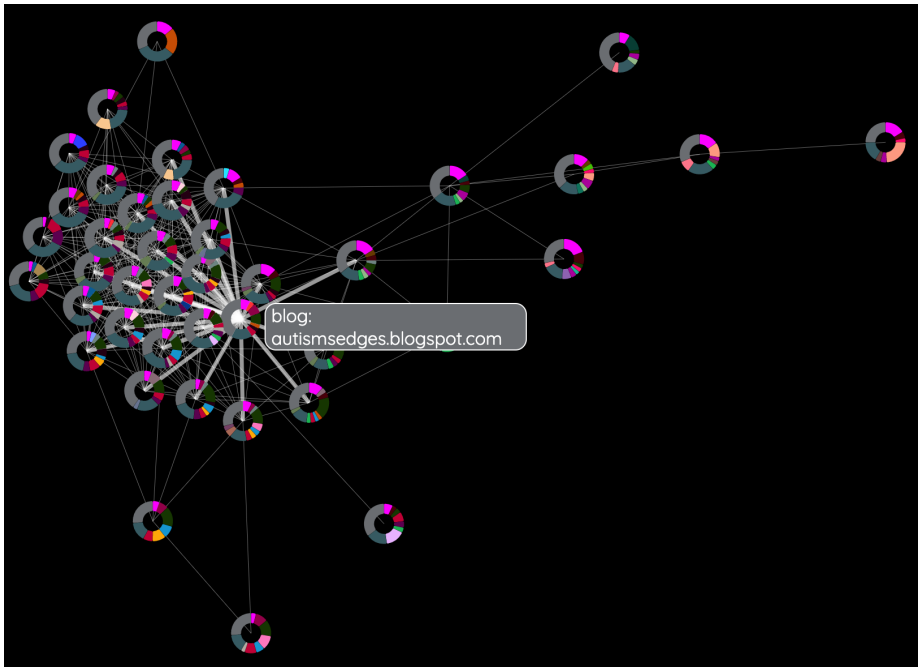

FIGURE 5. Force directed layout showing the proportions of each blog that pertain to each topic. Cosine similarity of these average topic proportions is used to determine the connections between blogs.

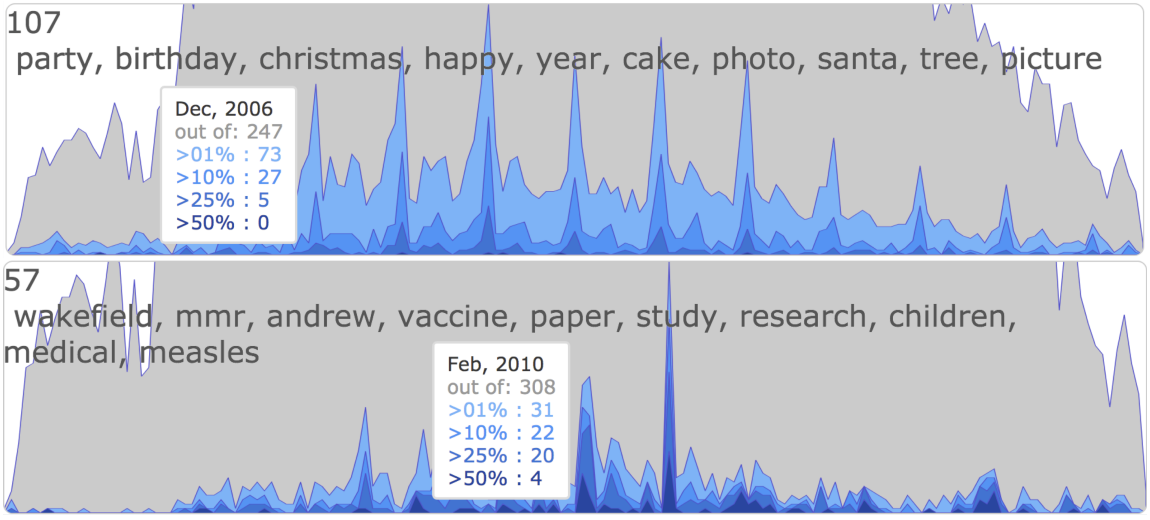

FIGURE 6. Early prototypes of a view showing changes in two topics' prevalence over time. This view was ultimately not included in Topicalizer, as it did little to advance the interpretivist research process.

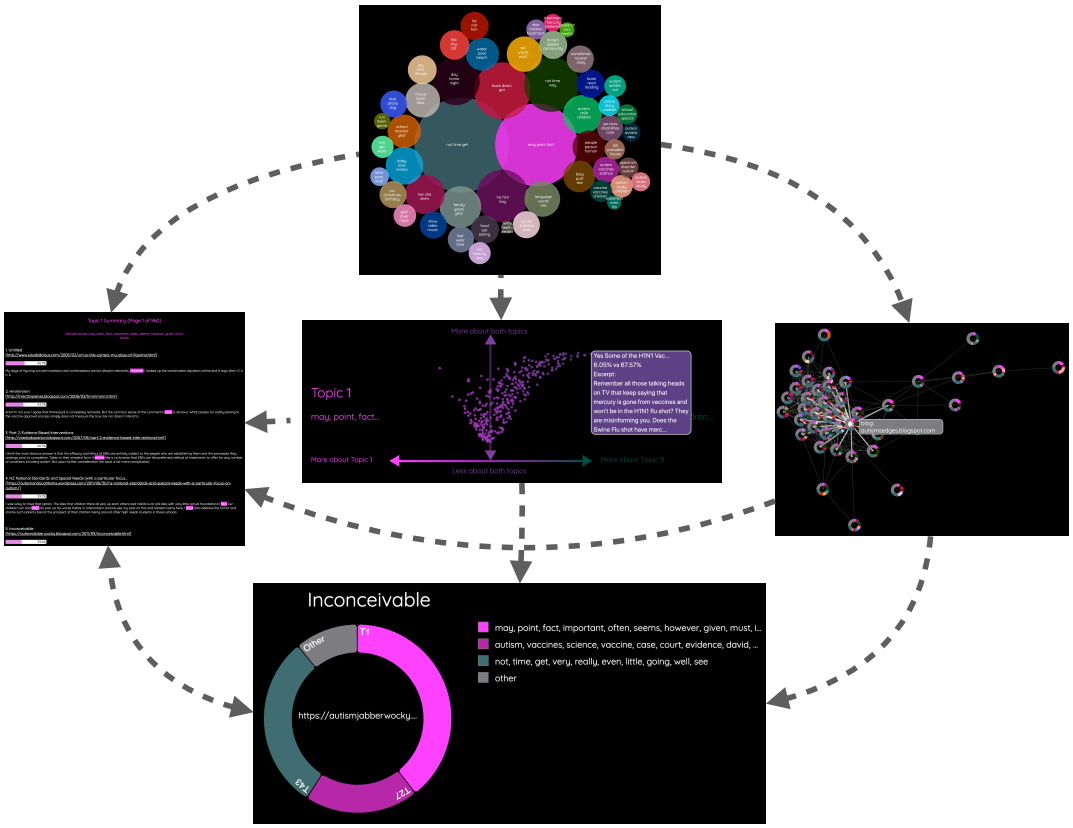

FIGURE 7. Flow chart of the different views in Topicalizer. Grey dashed arrows indicate how the user can transition among the different views by clicking on elements of the visualization. The visualization also provides a Back button, as well as the ability to return to the Venn diagram corpus view (at top in the above Figure) at any time.

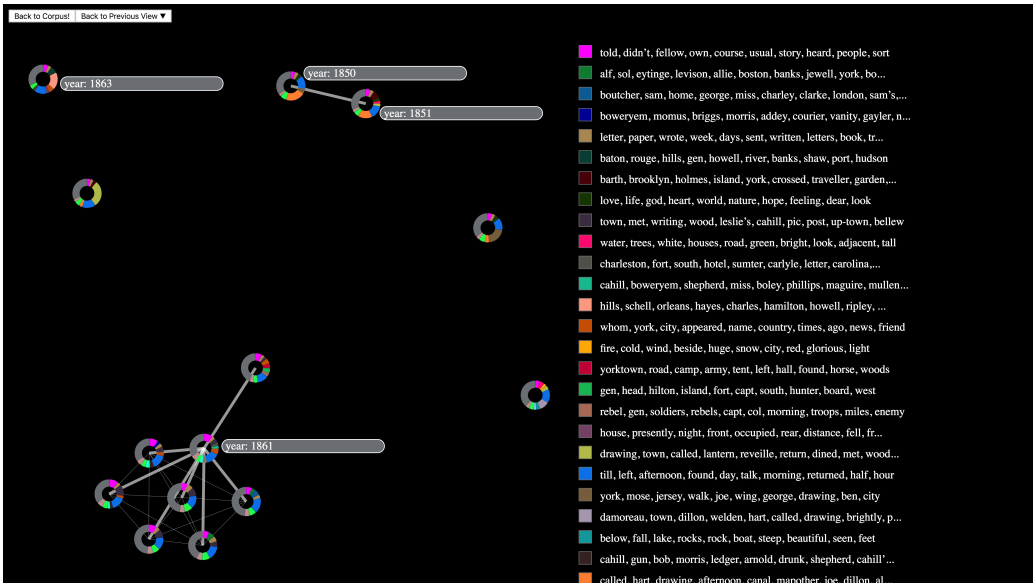

FIGURE 8. Composite screen shot of Gunn's diaries, aggregated by year. The year 1861 serves as a bridge, similar both to Gunn's writing in the late 1850s (clique near lower-left) and his coverage of the American civil war in 1862 (clique outlier).