The Influence of Process Management: Uncovering the Impact of Real-Time Managerial Interventions via a Topic Modeling Approach

Joshua T. Gyory
Mem. ASME
Department of Mechanical Engineering,
Carnegie Mellon University,
Pittsburgh, PA 15213
e-mail: jgyory@andrew.cmu.edu

Kenneth Kotovsky
Department of Psychology,
Carnegie Mellon University,
Pittsburgh, PA 15213
e-mail: kotovsky@cmu.edu

Jonathan Cagan[1]
Fellow ASME
Department of Mechanical Engineering,
Carnegie Mellon University,
Pittsburgh, PA 15213
e-mail: cagan@cmu.edu

Computationally studying team discourse can provide valuable, real-time insights into the state of design teams and design cognition during problem-solving. The particular experimental design, adopted from previous work by the authors, places one of the design team conditions under the guidance of a human process manager. In that work, teams under this process management outperformed the unmanaged teams in terms of their design performance. This opens the opportunity to not only model design discourse during problem-solving, but more critically, to explore process manager interventions and their impact on design cognition. Utilizing this experimental framework, a topic model is trained on the discourse of human designers of both managed and unmanaged teams collaboratively solving a conceptual engineering design task. Results show that the two team conditions significantly differ in a number of the extracted topics and, in particular, those topics that most pertain to the manager interventions. A dynamic look during the design process reveals that the largest differences between the managed and unmanaged teams occur during the latter half of problem-solving. Furthermore, a before and after analysis of the topic-motivated interventions reveals that the process manager interventions significantly shift the topic mixture of the team members' discourse immediately after intervening. Taken together, these results from this work not only corroborate the effect of the process manager interventions on design team discourse and cognition but provide promise for the computational detection and facilitation of design interventions based on real-time, discourse data. [DOI: 10.1115/1.4050748]

Keywords: cognitive-based design, design representation, design teams, topic modeling, process management

1 Introduction and Motivation

Design cognition studies investigate how designers think and strategize during design activities [1]. A variety of methodologies exist to analyze these processes of thought including, but not limited to, case studies, protocol analysis, and empirical performance tests [2,3]. Concurrent verbalization, or thinking aloud, also reveals certain characteristics of design cognition, such as the interactions between design problem and design solution [4]. By analyzing and codifying team discourse data, Stempfle and Badke Schaub proposed a two-process theory of thinking in design teams [5]. Their work theorizes that four basic cognitive processes encapsulate design thinking, including the operations of generation, exploration, comparison, and selection. More computational approaches have also been utilized to study design team communication, including Latent Semantic Analysis (LSA). LSA has been shown to be effective in modeling mental knowledge representations by analyzing design team communications such as emails, reports, and automating the annotation and tagging processes of team discourse to predict performance [6–10]. As the aforementioned works demonstrate, studying design team communication comprises a rich area. More critically, analyzing team interactions and communication amongst designers provide valuable insight into design processes and cognition, as well as predicting the state and effectiveness of problem-solving. Knowledge of this state/cognition of designers can provide feedback into identifying the types of facilitation and mediations teams may require during problem-solving.

Apart from design communication, a large body of work in engineering design theory research studies methodologies that facilitate problem-solving effectiveness. These include providing near and far analogies, patents, example solutions, and/or functional structures as inspirational stimuli, among others [11–15]. Along this vein, more recent work furthers these findings by studying the effects of intervening with design stimuli in near *real time*. Work by Goucher-Lambert et al. modulates the distance of design stimuli from designers' current design state *during* concept generation [16]. Midway through problem-solving, designers provide their current solution to a design problem. Then, a stimulus is computationally provided in real time to support ideation, either near to or far from the current state of the designer. In that work, Latent Semantic Analysis (LSA), a method utilizing the co-occurrence of words and singular value decomposition to represent the semantic space, computes the semantic similarity between current design states and a design space corpus of potential stimuli. Furthermore, work by Gyory et al. explores the effects of adaptive process management on design problem-solving [17,18]. Process managers guide design teams through problem-solving and intervene in *real time*, when deemed appropriate. That work shows that teams under the guidance of process managers significantly outperform

[1]Corresponding author.

Contributed by the Design Theory and Methodology Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received October 11, 2020; final manuscript received March 17, 2021; published online May 21, 2021. Assoc. Editor: Scarlett Miller.

teams that are not, in terms of their final design solutions. While both of these works move toward real-time, adaptive interventions during design problem-solving neither utilize discourse information in an algorithmic way to either monitor design progress or measure stimulus effects on design cognition. This research builds upon these studies (in fact, using data from the latter) and utilizes topic modeling to computationally analyze team discourse and detect the effects of real-time, process management.

Topic modeling spans a wide variety of algorithms, applications, and research domains. At the highest level, topic modeling constitutes a text mining process of automatically extracting themes or *topics* (which are sometimes latent) from corpuses of text [19]. By "latent," occasionally these themes/topics are not directly observable via direct inspection but are distinguishable through more nuanced and underlying similarities within the text. A specific type of method, used in this work, probabilistic topic modeling has been shown to be effective in identifying themes in unstructured text corpuses [20]. In all cases, the number of topics, specified by the researcher as an input to the algorithm, is assumed to be distributed across the entire corpus. In this way, the text corpuses can be modeled purely by their distributions over these topics. As an example, Rosen-Zvi et al. create an author-topic model from a collection of 1700 conference proceedings and 160,000 abstracts and illustrate its predictive power by revealing relationships between authors, documents, topics, and words [21]. Similar analyses have been done on other large collections of documents, such as in articles from Science [22].

Topic modeling has emerged as a valuable tool for researchers in the fields of engineering and engineering design theory. In transportation analysis, models are trained on corpuses of journal articles to identify sub-fields and provide a more holistic perspective on the current research landscape. In design research, Ball and Lewis model the expertise of members within multi-disciplinary design teams to predict their success and performance as teams in mass collaboration efforts [23]. Further examples in engineering design utilize topics and sentiment analysis to study design spaces [24]. Specifically, a Bisociative Information Network, composed of conceptual similarities between design topics, represents inspiration in idea generation as conceptual bridges between domain topics [25]. Additionally, in requirements engineering, Bhowmik et al. leverage topic modeling, augmented with part of speech tagging, to automatically generate requirements from stakeholders' comments to support combinatorial creativity. Joung and Kim analyze product designs and attributions through customers' perspectives through online reviews by using Latent Dirichlet Allocation and other semantic methods [26–28]. While topic modeling has emerged in the field of engineering, thus far, it has not seen much utilization to dynamically study design team cognition with the impact of process manager interventions, which offers promising opportunities to team research in engineering design, as explored in this work.

Consequently, this research utilizes a topic modeling framework to model discourse between members of a design team to study the impact of process manager interventions during problem-solving. Given that discourse provides insights into designer cognition and the state of the team over the course of problem-solving, the specific goal is to explore and computationally detect the impact of these interventions during the problem-solving process. To this end, Sec. 2 of this paper presents the cognitive design study from which the discourse data were collected in prior work [17], as well as a brief introduction to the field of topic modeling and the topic modeling framework for processing the verbalization data. Section 3 follows with results on the overall difference in topic structures between managed and unmanaged design teams, including a dynamic look at these topic structures over time. The second main analysis studies the direct impact of the interventions on team cognition, by focusing on and comparing the team members' discourse immediately prior to and following an intervention. Sections 4 and 5 conclude with a discussion of the results, particularly regarding process management and team discourse for engineering design teams, closing with opportunities for future research regarding the extension to real-time monitoring and intervention for design teams.

2 Background

This section provides background for two relevant areas of this work: Sec. 2.1 provides an overview of the cognitive study from which the transcript data originates, including the different experimental team conditions, the design prompt, types of design interventions, and data collection. The subsequent sub-section (Sec. 2.2) introduces the topic modeling framework for intervention assessment, discussing the core text analysis method selected as an illustration in this work, including the bag-of-words model and latent Dirichlet allocation (LDA).

2.1 Initial Motivating Study—Problem-Solving Under Process Management. The verbalization data presented in this work was collected from a behavioral study run with undergraduate, engineering students at Carnegie Mellon University. While only a general overview of the experiment methodology will be presented here, a more comprehensive outline is discussed by Gyory et al. 2019 [17]. Student designers were randomly placed into teams and allowed 30 min to solve the following engineering design problem:

Problem Statement:

Design a low-cost and easy to manufacture device that removes the outer shell from a peanut.

Constraint 1:

The device is meant to be utilized in developing countries where electricity may not necessarily be available as a power source.

Constraint 2:

In addition to the previous constraint, the proposed design must be able to separate a large quantity of peanuts from their shells, while causing minimal damage to the inner peanut.

The two constraints were dynamically added during problem-solving (10 min and 20 min into the experiment, respectively), to both exacerbate the difficulty of the problem and better emulate a dynamically evolving design task.

The experiment included three distinct team conditions, two of which are relevant to this study: managed teams and unmanaged teams. The managed teams comprised four student designers, all actively engaged in the problem-solving process and under the guidance of a human process manager (experienced, mechanical engineering graduate students). In this condition, the process managers intervened with their design teams, when they deemed it appropriate, in order to facilitate problem-solving. The interventions, described briefly next, were standardized across all managers. On the other hand, the unmanaged teams consisted of five student designers, all actively engaged in the problem-solving process and under the direction of a passive experimenter. These passive experimenters could only read instructions, provide the design prompt and constraints, and answer questions prior to the start of the experiment; otherwise, no communication between these facilitators and participants was permitted. The reduction in the number of student designers in the managed condition was meant to equalize the problem-solving resources between the managed (four active participants + manager) and unmanaged (five active participants) design teams.

The process managers in the managed team condition intervened with their design teams to affect the problem-solving process. While free to intervene when they deemed it necessary, the managers could only select interventions from a pre-determined *manager bank*. Their interventions were limited to this set. The process managers were also not allowed to speak with their design teams other

than to answer questions related to the experiment. This manager bank contained six design keywords, six design components, and six design strategies, all motivated by pre-existing strategies and techniques for increasing ideation/problem-solving effectiveness [29–32]. The design keywords and design components were related to the specific design prompt (e.g., "*sieve*" as a design component and "*high throughput*" as a design keyword), while the design strategies were tactics appropriate and generalizable to any design scenario (e.g., "Are the requirements being met in your current design? Can you go back to a previous idea?").

An audio recorder collected the design team discourse throughout the experiment. Sixteen transcripts were collected in total, with eight teams in the managed condition and eight teams in the unmanaged condition. The transcripts were transcribed via an outsourced vendor and manually checked for proper speaker identification. In addition to the transcripts, other data collected from the study included design teams' final designs (both a sketched drawing and a two-minute verbal description) and a complete recount of manager interventions, including the timing and type. These were actively noted by the managers during the experiment.

In the immediate days following the study, the researchers conducted a post-study interview with each of the process managers. These post-study interviews queried the managers on the motivations for and the perceived effects of their interventions ("What made you intervene with intervention [X]?" and "What was the effect of your interaction?"). The most common motivations noted include the following: trying to get all team members to equally contribute to the process; reminding the team of the constraints, requirements, and goals of the problem; and pushing teams to focus on a functional topic which they were either far away from our close to and needed an extra push, or to get the team back on track because they strayed completely away from the task [33]. These latter two motivations serve as the overarching inspiration for utilizing topic modeling as the computational approach in this work. If the managers felt that teams strayed from appropriate or conducive topics for effective problem-solving, can these topic shifts be computationally detected? Thus, this research addresses this question via a topic modeling framework to compare topic structures in the teams' discourse.

2.2 Text Analysis—Latent Dirichlet Allocation. While a variety of methods for semantic text analysis and topic modeling exist and can be used within the framework introduced in this work, the algorithm chosen for use in this paper is LDA with a unigram, bag-of-words assumption [34]. As both a generative and probabilistic method, LDA models a corpus of documents as a collection of underlying topics. No knowledge of these topics necessarily exists a priori, as the training procedure tests across varying values. LDA then generates topics from the distribution of words and documents in the corpus and probabilistically determines the fit with each number of topics. The researcher determines the number of topics and topic models with the best fit to the data and research goals, as presented in this paper in Sec. 4.1. Other methods for topic modeling include Latent Semantic Indexing (LSI), probabilistic latent semantic indexing (pLSI), non-negative matrix factorization (NMF), and Term Frequency—Inverse Document Frequency (TF-IDF) [35–37]. However, LDA overcomes some of the shortcomings of these precursor methods. Furthermore, LDA has been widely used on corpuses of many sizes, from micro-tweets, tweets, and micro-blogs, up to complete journal repositories [38–40], and results in more descriptive output, namely the actual topics that can be interpretable, and more differentiation to allow for potential explainability into the impact of the manager interventions. Bayesian probabilistic models, similar to LDA, have also been applied to discourse analysis [41,42]. Consequently, LDA was chosen for utilization in this work.

LDA assumes a bag-of-words model representation of the text corpus, one of the most commonly utilized methods in text analysis, information retrieval, and topic modeling [43]. Bag-of-words model captures the frequency of terms or words in each document and across the entire corpus of documents. The model in this work treats each individual word as its own feature (i.e., unigram model), though features can be represented by a sequence of n words through an n-gram model which retains information about word ordering [44]. The bag-of-words model assumes that the order in which these features appear in the text does not matter. In text analysis, a *word* represents the most basic unit of vocabulary and quantified by a basis unit vector: $\{1, \ldots, V\}$. In this case, a word signifies a unique utterance by a team member. It is from the co-occurrence of these words that underlying semantic similarities and topics can be inferred. *Topics* are defined as probability distributions over a set of words. *Documents* represent sequences of the words in the vocabulary, as denoted by $\mathbf{w} = (w_1, w_2, , w_3, \ldots, w_N,)$ for N words in each document, and the documents are generated by random mixtures over these topics, with topic mixtures, $(\theta_1, \ldots, \theta_M)$. The entire *corpus*, D, defines the entire collection of these M documents, $D = \{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \ldots, \mathbf{w}_M\}$.

For LDA's generative process, a Dirichlet distribution is used to represent the topic-word distributions across the documents, given by Eq. (1)

$$p(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \tag{1}$$

with K topics and concentration parameter α. The multivariate Beta function, B, normalizes this probability distribution. Then for each of the N words in each document, a topic is chosen, z_n, from a multinominal distribution with parameter, θ, and a word, w_n, is chosen from a multinominal probability conditioned on the topic z_n, $p(w_n|z_n, \beta)$. Finally, the joint distribution of a topic mixture, given the above parameters, is

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{K} p(z_n|\theta) p(w_n|z_n, \beta) \tag{2}$$

for a set of K topics, $\mathbf{z}$.

3 Methodology

A topic model first needs to be trained on the corpus of transcript data from the previously mentioned process management study. This section describes this training framework and sets up subsequent analyses on the output of the topic model. All of the natural language processing and text analysis algorithms throughout this work utilize MATLAB's implementation of LDA and supporting solvers.

3.1 Topic Modeling Framework. Figure 1 depicts the framework for training the topic model. The discourse of the design teams is identified by speaker, noting that the managed and unmanaged teams differ in the number of speakers (the unmanaged teams consist of five problem-solvers while the managed teams consist of four problem-solvers). The manager and experimenter dialogues are also removed from the discourse.

The transcripts are then further segmented into five-minute intervals to increase the corpus size per document, as needed to train sufficiently stable LDA models. In total, with eight managed and eight unmanaged teams, the entire corpus D consists of $M = 96$, distinct documents and $N = 741$ unique words. The average number of words per each five-minute interval document is 162 words, with a standard deviation of 25 words, and the average number of tokens per document being 452 tokens with a standard deviation of 102 tokens.

Prior to training the model, the transcripts undergo several pre-processing steps. The documents are first tokenized so that vectors of words represent each of the documents. All punctuations are removed, as well as the stop words identified in the Natural Language Toolkit (NLTK) [45]. Additional pre-processing steps

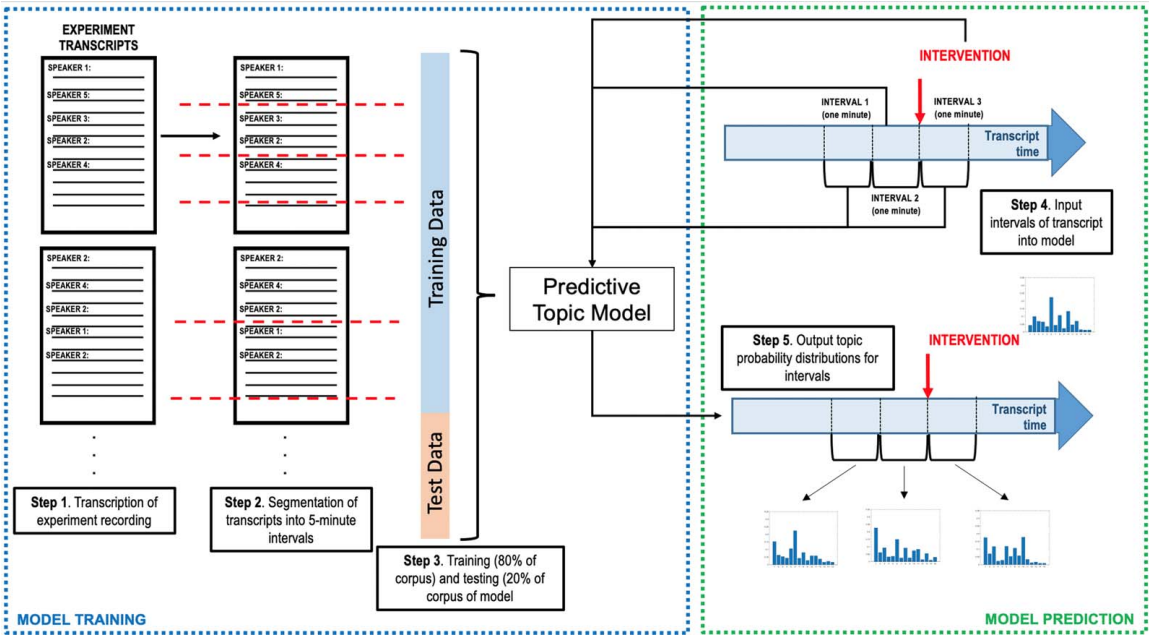

Fig. 1 Topic modeling framework with model training and model prediction

include the removal of infrequent words (those with frequencies less than 2), the removal of short (those less than two characters) words to eliminate noise with articles and non-words, as well as the stemming and lemmatization of words. These two processes remove prefixes and suffixes of the vocabulary, so that all words remain in the same tense and return to their dictionary forms.

During the topic model training, the corpus D is randomly split into a training set and a test set. The training set incorporates 80% of the corpus (77 documents), while the test set contains the remaining 20% of the corpus (19 documents). All of the five-minute interval documents are considered during the randomization of splitting between these sets, while the two added constraints during the experiment (recall the first was added 10 min in with the second added 20 min) could add some noise during training since the model is trained and testing on multiple randomizations of these two sets, and significant confounding effects from these constraints should be mitigated. Because the number of topics is not determined a priori for LDA, the topic model needs to be trained across a varying number of topics. Then, the number of topics with a better fit to the test data can be chosen. In this work, validation perplexity is used as one of the measures testing the fit to the data. Perplexity, a metric describing the goodness-of-fit, indicates how well the model, including the number of extracted topics, represents the documents; the lower the perplexity indicates a better fit [46–48]. The perplexity is mathematically defined in Eq. (3) as

$$perplexity(D_{test}) = \exp\left\{ -\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d} \right\} \tag{3}$$

with N_d being the total number of words and M being the number of documents.

3.2 Comparing Topic Mixture Outputs. Following the training of the topic model, the model can be leveraged to transform topic mixtures of smaller discourse segments within the transcripts. The topic model dimensionally reduces documents into a probability distribution over the topic space (i.e., topic mixtures). As mentioned in the prior study, process managers intervened with a bank of prescribed stimuli which included functional design components (hammer, conveyor, etc.) and design keywords (high throughput, sieve, etc.). As noted by the managers, these interventions were injected to direct team discourse. Therefore, comparing the discourse intervals immediately prior to and following an intervention should enable the detection of the impact of that intervention. As shown in Fig. 1, three different intervals are used to analyze the impact of the interventions, where I_t designates the time of the intervention. Interval 1 (I_{t-2}) includes the discourse between 2 min and 1 min prior to a process manager intervention, Interval 2 (I_{t-1}) includes team communication between 1 min prior to and immediately up to the point of the intervention, and Interval 3 (I_{t+1}) includes the time immediately following the intervention and up to 1min after. An assumption of this work defines an "effective" topic shift as to whether or not an intervention causes a topic shift *toward* the topic mixture of the intervention itself, indicating converging discourse on the intervention topic. As an illustrative example, an allowable design component intervention includes a sieve. The manager may intervene with the sieve component to get their team to start focusing on sorting the peanuts from the crushed shells. The goal is to computationally uncover whether the team starts discussing not necessarily sieve specifically, but the concept, design, and/or function of sorting in general. Again, this idea of shifts in topic originates from the post-study interviews conducted following the behavioral study, in which topic-related motivations emerged as a common theme throughout the manager interventions. The immediate goal is to determine whether those motivations are realized in producing effective behavior change. This overall notion motivated the inspection of Interval 2 and Interval 3. Interval 1 is included in the analysis as a control, controlling for the idea that teams were already moving closer to the intervention topic. These analyses are described in more detail later in the paper.

To fully carry out the above analysis, a topic mixture for the intervention itself must be defined. Recall that the design keywords and design function interventions consist of a single word and/or image (e.g., the text and image of a *sieve* for a design component, and the text of *high throughput* for the design keyword). Consequently, in order to define the intervention topic mixtures, the full dictionary definition of the words are used. For example, Merriam-Webster dictionary defines *sieve* as *a device with meshes or perforations through which finer particles of a mixture (as of ashes, flour, or sand) of various sizes may be passed to separate them from coarser ones, through which the liquid may be drained from liquid-containing material, or through which soft materials may be forced for reduction to fine particles* [49]. On the assumption that the

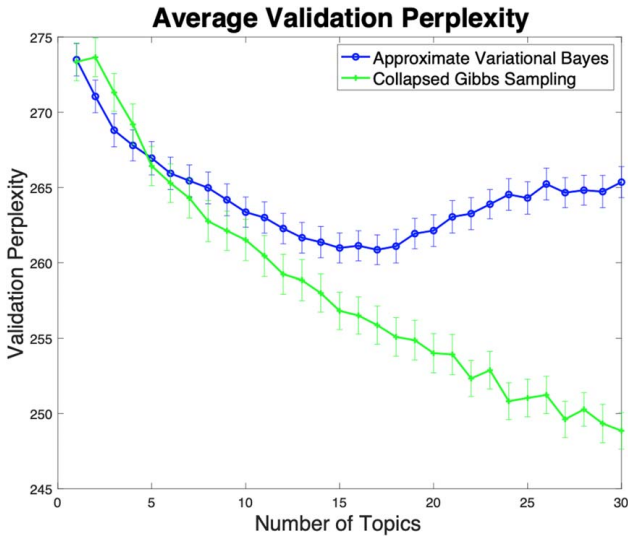

Fig. 2 Average validation perplexity for two LDA algorithms over a varying number of topics (error bars show ±1 S.E.)

words in the definition are relevant to the term, this text document for the intervention, when mapped into the topic space using the trained model, can now be used to define the topic mixtures of the interventions. In this above-mentioned example, an intervention is perceived as effective not only if the team starts talking specifically about a sieve, but if they also start discussing meshes, the concept/function of separating, etc. Thus, including the dictionary definition as the topic mixture for the interventions provides this additional level of detail for detection. The intervention documents are not included in the training of the topic model itself.

After extracting the topic probability distributions from the topic model, the distributions from the different intervals and the intervention need to be compared with each other. To measure the similarity between the topic probability distributions, the Kullback-Leibler (KL) divergence is used, also known as information divergence or relative entropy [50,51]. The KL divergence computes how different one probability distribution is from another probability distribution. It is more precisely defined in Eq. (4), for a discrete probability distribution, as

$$D(P\|Q) = \sum_{\mathbb{R}^K} p(x) \log \frac{p(x)}{q(x)} dx \tag{4}$$

where P and Q represent two discrete, probability distributions, over the same variable, x (which in this case are the topics). As P and Q approach one another in similarity, the KL divergence approaches zero. It should be noted that the order for KL divergence matters, as the metric is not symmetric and does not follow the triangle inequality [52,53]. In other words, the divergence from P to Q does not necessarily equal the divergence from Q to P. Thus, in order to make relative comparisons, all analyses in this work follow the same temporal ordering (relative to the timeline of the experiment) for the distributions.

4 Results

Section 3.1 first discusses the selection of the topic model with a specific number of topics. This includes results from different optimization solvers, perplexity, pointwise mutual information, as well as a parametric analysis on the concentration priors that describe the prior word and topic distributions. Next, analysis on the differences in design cognition, modeled as topic mixtures, between the managed and unmanaged teams is presented both statically (the overall transcripts themselves) and over time (Sec. 3.2). Then, a before and after analysis of specific manager interventions on the discourse detects the impact of the process managers on design cognition via directed topic shifts (Sec. 3.3).

4.1 Topic Model—Validation and Selection. As discussed in the previous section, the number of topics for topics models is not determined a prior, so the model must be trained over a varying number of topics. The model is trained across a range of one to thirty topics, and for each distinct number of topics, trained for 100 iterations. Each iteration randomly selects the training and test sets from the corpus. Recall that the training set consists of 80% of the corpus, or approximately 77 documents, while the test set consists of the remaining 20% percent of the corpus, or 19 documents. Figure 2 shows the average validation perplexity of the 100 iterations over the range of topics: from one to thirty. Different optimization solvers are used in the fitting of the model to test performance. These include stochastic approximate variational Bayes [54,55], collapsed Gibbs sampling [56], approximate variational Bayes [57], and zeroth order, collapsed variational Bayes [57,58]. While more in-depth comparisons between the four solvers lie outside the scope of this work, here the different solvers are compared by their fit to the data via validation perplexity on the held-out test set.

Figure 2 shows validation perplexity for the two better-performing optimization solvers: approximate variational Bayes and collapsed Gibbs sampling. The lower the perplexity indicates a better fit. The other two solvers result in significantly worse performance, with stochastic variational Bayes showing no increase in performance with an increasing number of topics and thus omitted from the figure. The two solvers shown behave a bit differently. While the validation perplexity decreases with an increasing number of topics for both, collapsed Gibbs sampling continues to decrease while approximate variational Bayes reaches a plateau before starting to increase in validation perplexity. While approximate variational Bayes plateaus at around 15 topics, collapsed variational Bayes starts to become noisy around 20 topics and greater. Due to this difference in behavior, both are considered in choosing the number of topics as well as an additional measure—pointwise mutual information.

Pointwise mutual information (PMI), or topic coherence, is measured across the same range of topics as the perplexity. PMI is occasionally used instead of, or in conjunction with, perplexity to characterize topic models, as it is been shown that perplexity can negatively correlate with human perception of the topics [59]. While human judges can be used, previous work in this area has demonstrated that it is possible to automatically measure topic coherence with near-human accuracy using topic coherence [60,61]. Specifically, pointwise mutual information scores the probability of pairs of terms taken from topics and their appearance across topics. In other words, PMI identifies overlap of information contained in topics. The same two better-performing algorithms (in terms of validation perplexity) are graphed in Fig. 3, which shows the average normalized PMI, calculated on the corpus, across the topics and models. As shown in the figure, both algorithms see a similar trend in pointwise mutual information gains. Both experience significant increases early on with a smaller number of topics, both starting to settle between 12 and 18 topics. These trends are mutually considered with the previous trends from the validation perplexity to characterize and choose a topic model.

Taken together, both the validation perplexity and pointwise mutual information measurements support a similar range of topics. Since there are no significant differences in validation perplexity over the range from 15 to 20 topics (the two solvers are within a range of 4.0), and both solvers' PMI level around a similar range, 15 topics (the plateau in perplexity) is chosen throughout the remainder of this work and analysis. Now regarding the specific model, recall that 100 different topic models are averaged for each number of topics. For those models using 15 topics, the model with the lowest validation perplexity, across both of the solvers, is selected. The model chosen possesses a

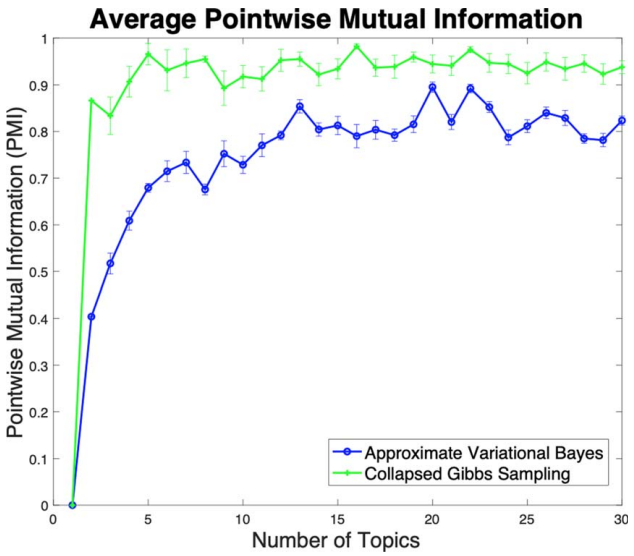

Fig. 3 Average pointwise mutual information for two LDA algorithms over a varying number of topics (error bars show ±1 S.E.)

validation perplexity on the test set of $p(D_{test}) = 219.53$. As a final level of validation, the cosine similarity is computed between all pairs of resulting topics. Measuring the intra-topic similarity between topics provides an additional measure of overlap between topics—the greater the similarity and overlap, the less distinct the topics are. Ideally, topics should not significantly overlap. Excluding self-similarities, the average pairwise cosine similarity between topics for the selected model is $\mu_{c.s.} = 0.06$.

After selecting a specific topic model, a parametric analysis is performed to identify the sensitivity of this model with varying values of the two hyperparameters. The two hyperparameters for LDA, α and β, describe the prior distributions on the topic and word concentrations, respectively, when fitting the model. Larger values of α define documents as being composed of a wider variety of topics while larger values of β define topics as being composed of a wider variety of words. Figure 4 shows the resulting surface from the parametric analysis on the hyperparameter priors. For the analysis, values of the priors vary from $\alpha, \beta \in [0.01–1]$ in 100 equal intervals. As shown, with increasing values of both α and β, the validation perplexity value decreases. Only with low values of the

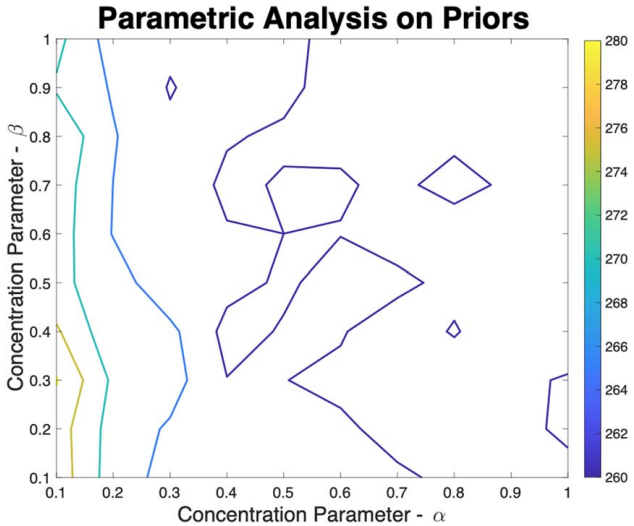

Fig. 4 Parametric analysis on validation perplexity, varying prior parameters α and β

hyperparameters, specifically at approximately $\alpha \leq 0.4$, does the result become sensitive, with a significant and steep increase in validation perplexity. Accordingly, the chosen model uses values of these hyperparameters with lower and less sensitive perplexity.

As defined earlier, the topics are represented by a vector, equal in length to the number of words in the entire corpus. The entire document corpus D, from the 16 transcripts, contains $N = 741$ distinct words after the aforementioned pre-processing steps and thus represents the size of each topic vector. Each word in these topic vectors is assigned a probability value that appears in that particular topic. Therefore, these topics can be "visualized" by viewing the most probable words in the respective topics. Table 1 shows the ten most probable words for each of the 15 resulting topics from the chosen topic model. Additionally, to illustrate a sample distribution within a topic, Table 2 shows the ten most probable words for Topic 11, along with their associated probabilities of occurrence.

4.2 Comparing Topic Mixtures Between Managed and Unmanaged Teams. After the topic model has been sufficiently trained, the model can be used to transform different documents into the topic space. This dimensionality reduction of discourse using the trained model produces a topic mixture profile for the input document. The topic mixture shows the probability for all of the fifteen topics appearing in that specific document. This initial analysis shows the topic mixtures of the entire transcripts themselves. That is, each of the eight managed and unmanaged teams' transcripts is projected into the topic space, and the results are shown in Fig. 5. Each bar represents the average probability of the topic for both the unmanaged and managed teams. For example, in terms of Topic 1, the average probability that this topic appears across the managed teams' discourse is 0.053, or 5.3%, while the average probability across the unmanaged teams' discourse is 0.12, or 12%.

Across the 15 different topics, using a two-tailed, non-parametric Mann Whitney U-test, four topics exhibit significant differences between the two team conditions: Topic 1 ($p < 0.0047$), Topic 10 ($p < 0.028$), Topic 11 ($p < 0.05$), and Topic 12 ($p < 0.027$). Topics 10 and 11 highlight an important finding from this analysis, which exhibit significantly higher probabilities of appearing in the managed teams' discourse. These topics, visualized by their ten most probable terms (Table 1), contain more design components, for example, sieve, funnel, belt, and roller, with some of these functional components coming directly from the manager interventions. Thus, the managed teams are significantly more likely to discuss these functional components and directly influenced by the manager interventions, which contain such functional concepts.

In addition to mapping the entirety of the transcripts into the topic space, five-minute discourse intervals can also be mapped (i.e., $t \in \{0, 5\}$, $t \in \{5, 10\}$, $t \in \{10, 15\}$, $t \in \{15, 20\}$, $t \in \{20, 25\}$, $t \in \{25, 30\}$, where $t \in \{x, y\}$ defines the interval from x minutes to y minutes in the experiment). Transforming to these smaller intervals provides a dynamic look at the topic distributions over the course of problem-solving, as well as the differences between the managed and unmanaged teams over these time periods. Figure 6 shows the difference in probability distributions over the topic space for those six specific time intervals. A positive difference indicates that the topic exhibits higher probability in the managed teams while a negative difference indicates higher probability in the unmanaged teams.

Overall, computing the sum of the squared differences (SSD), the second half of the experiment shows greater disparities between the discourse of the managed and unmanaged teams, with the final interval exhibiting the largest, as shown in Table 3.

This result shows that the process managers create a larger impact during the second half of the experiment, as these intervals contain the largest overall differences in team discourse. Again, utilizing a two-tailed, non-parametric Mann Whitney U-test, several significant differences in each of the interval's individual topics emerge

Table 1 The 15 extracted topics with the ten most probable words in each topic

TOPICS	TOP 10 WORDS
Topic 1 *(nPMI* = 0.664*)*	Side, good, wait, top, bottom, draw, large, middle, view, kinda
Topic 2 *(nPMI* = 0.862*)*	Peanut, easy, manufacture, cost, remove, low, guess, metal, electricity, plastic
Topic 3 *(nPMI* = 0.752*)*	Peanut, hole, sort, big, hard, guess, bit, split, talk, apply
Topic 4 *(nPMI* = 0.783*)*	Design, kind, open, idea, cut, thing, work, nutcracker, slide, start
Topic 5 *(nPMI* = 0.742*)*	Mmhmm, affirmative, time, push, crush, inside, constraint, work, fall, claw
Topic 6 *(nPMI* = 0.811*)*	Peanut, crack, kind, device, blade, size, roll, half, move, mechanism
Topic 7 *(nPMI* = 0.749*)*	draw, circle, guy, minute, sheet, long, thing, mumble, handle, suppose
Topic 8 *(nPMI* = 0.850*)*	Crush, small, piece, thing, separate, break, basically, pressure, force, move
Topic 9 *(nPMI* = 0.740*)*	Peanut, thing, hand, pull, press, clamp, spring, speaker, fall, hold
Topic 10 *(nPMI* = 0.686*)*	Sieve, nut, funnel, high, add, shake, good, simple, amount, person
Topic 11 *(nPMI* = 0.742*)*	Crank, conveyor, peanut, belt, fall, roller, final, turn, attach, leave
Topic 12 *(nPMI* = 0.778*)*	Feel, wood, pretty, fine, flat, cheap, happen, edge, stuff, easily
Topic 13 *(nPMI* = 0.819*)*	Basically, wheel, hmm, connect, power, process, spin, human, large, edge
Topic 14 *(nPMI* = 0.799*)*	Laughs, gear, lot, laugh, wire, pretty, box, gap, foot, version
Topic 15 *(nPMI* = 0.689*)*	Replaces, nut, true, rotate, draw, screw, wall, sense, cylinder, call

Table 2 Sample topic showing the ten most probable words and their associated probabilities

Word	Probability
Crank	0.122
Conveyor	0.063
Peanut	0.062
Belt	0.0621
Fall	0.0606
Roller	0.0484
Final	0.0469
Turn	0.0378
Attach	0.0363
Leave	0.032

between managed and unmanaged teams: 5 min (Topic 7 ($p < 0.01$) and Topic 8 ($p < 0.01$)), 10 min (Topic 9 ($p < 0.04$)), 15 min (Topic 3 ($p < 0.01$)), 20 min (Topic 1 ($p < 0.05$), Topic 10 ($p < 0.01$), and Topic 13 ($p < 0.01$)), 25 min (Topic 1 ($p < 0.01$), Topic 8 ($p < 0.05$), Topic 10 ($p < 0.02$), and Topic 12 ($p < 0.04$)), and 30 min (Topic 1 ($p < 0.001$) and Topic 12 ($p < 0.02$)). Topics 10 and 11, which both contain functional concepts, become increasingly more integral to the managed team members' discourse as problem-solving progresses, with Topic 10 becoming significantly more integral during the 20-min and 25-min intervals. On the other hand, Topic 1 becomes increasingly more integral to the unmanaged team members' discourse, and this trend continues throughout the entirety of the experiment, becoming most significant in the final interval. Visualizing Topic 1 from Table 1 (the most probable words in the document), the topic focuses more on

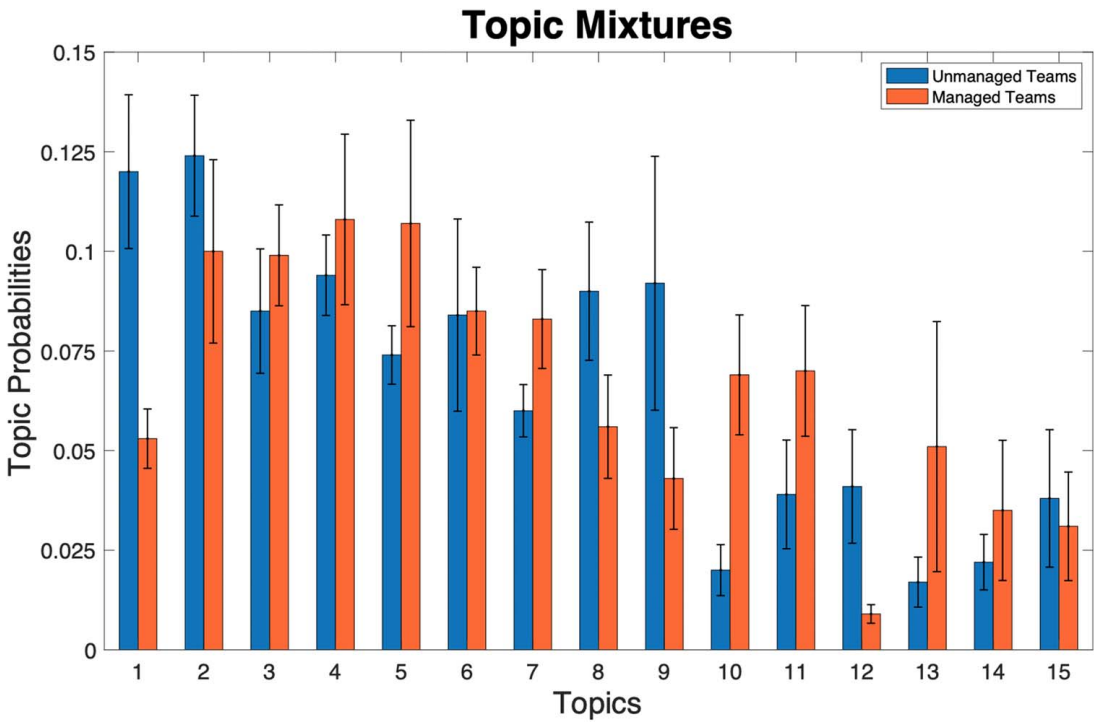

Fig. 5 Topic mixtures (topic probability distribution) for managed and unmanaged teams' transcripts (error bars show ±1 S.E.)

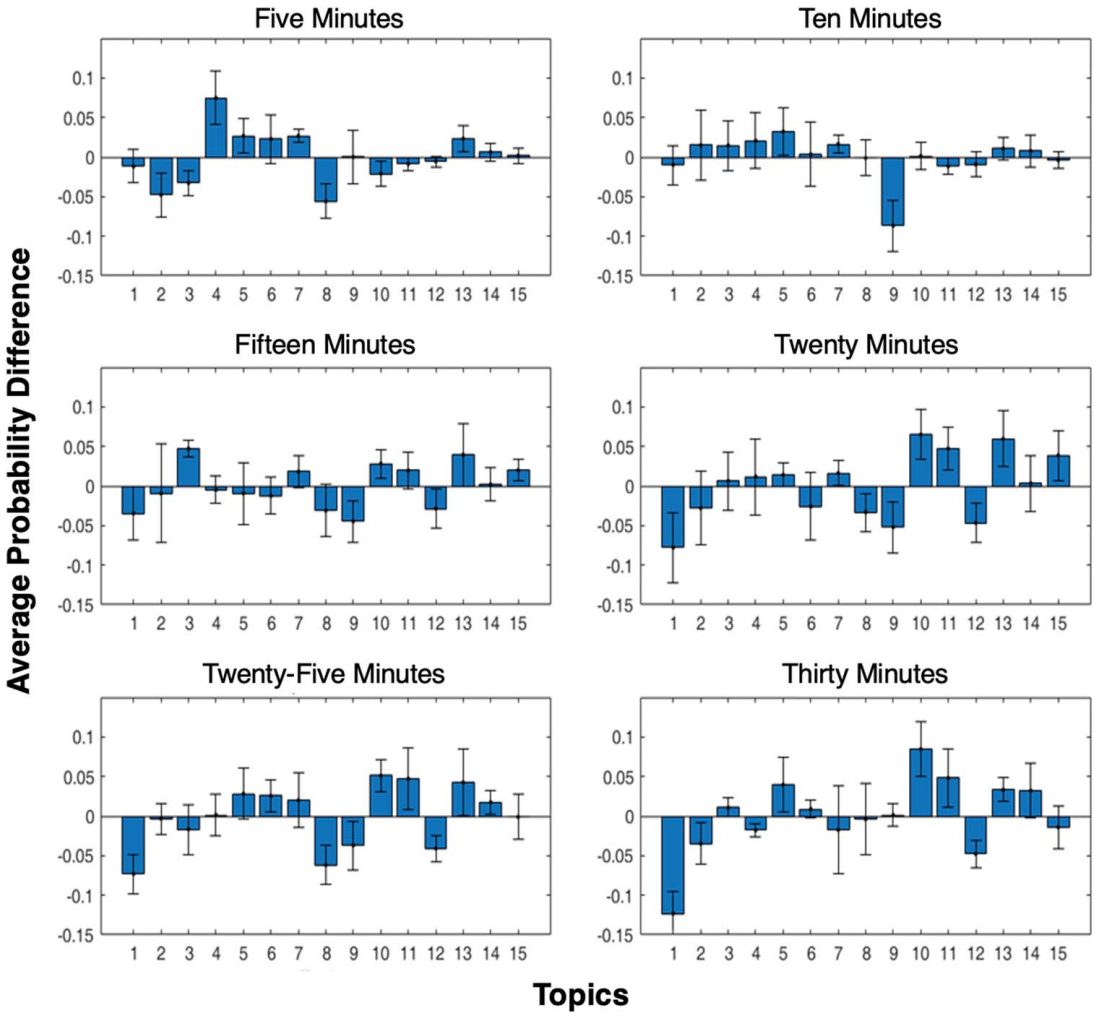

Fig. 6 Difference in topic mixtures (topic probability distributions) for managed and unmanaged teams over time (error bars show ±1 S.E.)

the abstract design process and structure rather than on concrete design functions, such as "draw," "top," "bottom," "middle," and "view." Linking this to the semantics of the design process, the discourse in this topic seems to be based on visualization and orientation, and contradicts the process manager strategies, whose focus near the end was to home their teams in on functional concepts and ideas. This significant focus, particularly at the end of the experiment when designs need to be past the abstract/conceptual stage, could have been one of the factors harming the unmanaged teams, ultimately leading them to their inferior performance. However, further analyses to isolate this effect would be needed to determine the extent to which this focus detrimentally impacted the unmanaged teams.

4.3 Detecting Effects of Manager Interventions. The next analysis focuses on the effects of individual manager interventions on team discourse. As depicted in Fig. 1, the topic mixture of the interval leading up to an intervention (I_{t-1}) is compared to the topic mixture of the interval immediately following an intervention (I_{t+1}). The assumption of this analysis relies on the idea that the team members' discourse should be more aligned with the intervention immediately following the manager intervention compared to before. For this analysis, 20 distinct interventions are studied, because (1) the intervention is either a *design keyword* or a *design component,* or (2) the intervention is the specific design strategy intervention of, "*Can you identify the assumptions, constraints, and goals of the problem*?", and (3) no other interruptions (i.e., constraints or manager interventions) occurred within the one minute prior to and following the intervention. The first two requirements ensure more topic-focused, concrete interventions. The remaining design strategies are more process-related and thus unsuitable for study via the topic modeling framework. Accordingly, future work can consider how to computational detect these remaining design strategy interventions. The third criteria (that no other interruptions occur within the one-minute intervals) control for other confounding variables in the analysis that could potentially cause additional topic shifts. As mentioned previously, these 20 interventions came from a broader set of 52 total process manager interventions.

Table 3 Squared differences between managed and unmanaged teams' discourse via the topic space

Time (t—min)	$t \in \{0, 5\}$	$t \in \{5, 10\}$	$t \in \{10, 15\}$	$t \in \{15, 20\}$	$t \in \{20, 25\}$	$t \in \{25, 30\}$
Squared difference (SSD)	SSD = 0.015	SSD = 0.010	SSD = 0.011	SSD = 0.026	SSD = 0.021	SSD = 0.033

For these 20 distinct interventions, the KL divergence computes the similarity between the topic probability distributions of the one-minute interval prior to the intervention (I_{t-1}) and the one-minute interval after the intervention (I_{t+1}), both against the topic probability distribution of the intervention itself. Utilizing a two-tailed, non-parametric Mann Whitney U-test, results show that the topic mixtures in the minute following the intervention are *significantly more similar* to the intervention topic mixture than prior to the intervention ($I_{t-1} = 0.69$, $I_{t+1} = 0.46$, $p < 0.005$, *effect* $r = 0.44$, and $U_{crit} = 127.5$). This result indicates that the interventions generate a significant impact on the topic structure and design cognition of the design teams, leading them to direct the focus of their discourse on the provided interventions.

To further corroborate this finding, an LSA model is also trained on the data. Similar in spirit to LDA, LSA instead utilizes singular value decomposition for dimension reduction as opposed to a probabilistic approach as LDA does. The semantic distances (via cosine distance, D_c) between the one-minute discourse intervals, both before and after the interventions, are computed with the intervention documents. Across the same 20 interventions, the average cosine distance prior to an intervention is $D_c = 0.67$ with the average cosine distance following the interventions is $D_c = 0.45$, with the average change (using a temporal change with respective to experimental time, i.e., before intervention minus after intervention) in distance being $\Delta D_c = 0.124$. Again, the results indicate that the discourse immediately after an injected intervention is more similar (smaller D_c) than prior to the intervention.

A follow-up analysis provides further evidence that this similarity in topic mixtures is a direct effect of the interventions. In order to rule out the possibility that the design teams incrementally move closer and closer to a given topic over time on their own, the KL divergence between all three intervals' topic mixtures with the intervention topic mixtures is computed. Thus, for the above-mentioned notion to hold, the divergence between the first interval's topic mixture and the intervention topic mixture should be the largest and then decrease through I_{t-1} and I_{t+1}. After computing the divergences, the changes between the before and after intervals of the intervention can be computed, as in Eqs. (5) and (6)

$$\Delta_{12} = D(\theta_{I_{t-2}} || \theta_{Inter}) - D(\theta_{I_{t-1}} || \theta_{Inter}) \tag{5}$$

$$\Delta_{23} = D(\theta_{I_{t-1}} || \theta_{Inter}) - D(\theta_{I_{t+1}} || \theta_{Inter}) \tag{6}$$

where $\theta_{I_{t-2}}$ is the topic mixture for I_{t-2}, $\theta_{I_{t-1}}$ is the topic mixture for I_{t-1}, $\theta_{I_{t+1}}$ is the topic mixture for I_{t+1}, θ_{Inter} is the topic mixture for the intervention, and D is the KL divergence operator. For example, the first term in Eq. (5) ($D(\theta_{I_{t-2}} || \theta_{Inter})$) computes the KL divergence from I_{t-2}'s topic mixture to the intervention's topic mixture.

Conceptually, Eq. (5) shows whether the team members' discourse becomes more similar or dissimilar to the intervention *prior to* the intervention, while Eq. (6) provides the same information, but *over* the intervention. Again, utilizing a two-tailed, non-parametric Mann Whitney U-test, results indicate that the changes between the intervals' topic mixtures from Eqs. (5) and (6) are significantly different ($\overline{\Delta}_{12} = -0.14$, $\overline{\Delta}_{23} = 0.23$, $p < 0.003$, *effect* $r = 0.48$, $U_{crit} = 113.4$). Not only are the average divergences significantly different, but they are opposite in sign. Consequently, in the two, one-minute intervals prior to the intervention, the team members' discourse drifts away from that of the intervention, (i.e., becomes more *dissimilar* to the intervention), while during the one-minute intervals before and after the intervention, the team members' discourse converges back to the intervention topic (i.e., becomes more *similar* to the intervention). Thus, the possibility that the design teams incrementally move closer and closer to a given topic over time is disproven, further corroborating that the topic shifts are caused directly by these manager interventions.

5 Discussion

The utilization of topic modeling for the framework in this paper, to detect the effect of the intervention on design team process, is motivated by several factors. The first, as mentioned previously, involves the post-study interviews conducted with the process managers. After querying the human managers on their rationale for intervening with their design teams, many consisted of topic-related rationales. For example, one manager mentioned, "*Uhh, they* [the team] *were getting really close to the idea and they had been really close to the idea of a blade for a long time, but it was just to push them a little bit farther in that direction. They had the drawing and were really close, but they started adding more complicated things that I did not really think would be helpful.*" Additionally, another manager indicated, "*They* [the team] *started talking, they were very close to the idea, wanting to have a sieve, but couldn't come up with the idea themselves. They were trying to think of much more complicated solutions for that.*" These quotes from the managers, representing just two of many, highlight the concept of topic pushes or topic shifts. Topic modeling provides an algorithmic way to computationally detect these changes and shifts in the topic. Accordingly, the overarching question this research answers is whether these topics shifts can be computationally detected and reveal the plausible mechanism underlying the effectiveness of the managerial interventions. An additional motivation for utilizing topic modeling to answer the aforementioned research question lies in the algorithm output. The output of LDA, the illustrative technique used in this paper, shows a probability distribution over a range of topics. Accordingly, this distribution provides a more holistic representation of the discourse data, as it maps team interactions along a spectrum of topics. This contrasts with other work in this area using manual/automatic coding, which maps lines of discourse to a single coding scheme.

The topic model goes through an exhaustive training procedure, varying over a number of topics, optimization solvers, and training and test sets. For each number of topics and optimization solver, the model runs for 100 iterations (each data point in Figs. 2 and 3 averages across those 100 runs). The results for the two better-performing optimization solvers, collapsed Gibbs sampling and approximate variational Bayes, show that the perplexity behaves a bit differently. While the validation perplexity decreases with an increasing number of topics for both, collapsed Gibbs sampling continues to decrease while approximate variational Bayes reaches a plateau at around 15 topics. Collapsed variational Bayes starts to become noisy at 20 topics and greater. In addition to perplexity, pointwise mutual information (PMI), or topic coherence, is measured across the same range of topics. The two better fitting solvers in terms of validation perplexity are graphed in Fig. 3, which shows the average PMI across the topics for a model. Both experience significant increases early on with a smaller number of topics and start to settle between 12 and 18 topics. For the selection of the number of topics, this two-pronged approach acknowledges the limitations of perplexity when it comes to the perceptibility of topics via direct human inspection. Perhaps not the absolute optimal model, taking all these factors into consideration results in a pragmatically sufficient topic model for further analyses. A parametric analysis also tests the sensitivity of the hyperparameters describing the prior distribution, and a cosine similarity metric tests the overlap in the resulting topics.

The goal of this work is not to compare the efficiencies or performance of different types of topic models and/or algorithms. Rather, the goal is to study the effects of managerial interventions via design discourse. Since there does not exist a ground truth for the discourse data (a prior knowledge of the exact topics to extract), it is difficult to formalize a precise measure of optimality or the "absolute best" model. The rigor of the model selection process for this work explores the space of models with LDA and selects the one that provides acceptable performance on the chosen metrics. Used consistently across conditions, the chosen model can then be used to analyze differences between teams and

segments of discourse. In addition, an LSA model further supports the overall findings in the convergence of discourse as an impact of the process manager interventions. Future work can explore how different modeling algorithms compare to LDA (besides LSA which already corroborates results in this work) and perform on the discourse data. An additional opportunity can also explore training on larger corpuses such as Wikipedia and Google News.

With the model trained, two different analyses test whether the interventions can be detected within the design team discourse. The first (Sec. 4.2) maps the entire transcripts, comparing the managed and unmanaged teams, along the topic space by outputting their associated topic mixtures (i.e., the topic probability distributions). Two interesting findings emerge. The first involves the topics most relevant to the design interventions, specifically Topic 10 and Topic 11. These two topics are significantly more probable in the managed teams, including more functional concepts and are more representative of the design component interventions. This result is validated by both (1) visualizing these two topics with their 10 most probable words and (2) mapping the interventions themselves into the topic space (where topics 10 and 11 become more prominent). The second interesting finding involves Topics 2–4. All three of these topics are nearly equal between the managed and unmanaged teams and, apart from topic one, contain the highest probabilities across the space. These topics most pertain to the constraints and goals of the problem, so it is not surprising (and further validates the framework) that these topics appear highly and equally probable across both team conditions.

A dynamic look at the topic mixtures across the experiment is also performed. Two interesting trends emerge from comparing the topic mixture space between the managed and unmanaged team conditions. Topics 10 and 11, which both contain more functional concepts, become increasingly more integral to the managed team members' discourse throughout the experiment, particularly near the end. Furthermore, Topic 1 becomes increasingly more integral to the unmanaged team members' discourse, reaching significance from the managed teams in the last two intervals of the experiment. The visualization of this topic shows more abstract, design process-related activity. Taken together, these two results point to managers guiding the solving process toward completion and refinement of the final designs, while the unmanaged teams tend to be more focused on design visualization. Future work can consider smaller intervals of the transcript, to gain more resolution in the evolution of the topic structures throughout problem-solving.

The second analysis (Sec. 4.3) performs a before and after investigation on 20 distinct manager interventions. Consequently, as shown in Fig. 1, the one-minute intervals prior to an intervention and immediately following an intervention are mapped into the topic space. Using KL divergence to compare the similarity of probability distributions, the topic mixtures of these transcript intervals are compared to the topic mixture of the interventions themselves. Results reveal that, on average, the discourse becomes significantly more similar to the intervention immediately after the intervention. Taken together, these findings validate the detection of the topic shifts in discourse, which many of the managers claimed as their motivation for intervening. Of these 20 interventions studied, in four instances, the topic mixtures of the interval following the intervention actually become more *dissimilar* to the topic mixture of the intervention. These cases deserve a more thorough investigation. One of these cases has a near-zero change, while one of these four cases has a significantly larger difference than the others. In this particular instance, the manager directly perceives the intervention as ineffective, saying that "It really did nothing, not at all." In this case, it is interesting to note consistency in how the manager perceives the intervention with the detection within the discourse, as the team members' topic mixtures become more dissimilar immediately after the intervention. In general, deeper dives into these outlying cases can provide additional insight when considering the implementation and effectiveness of a real-time, intervention framework.

Overall, this work shows promise in a more automated approach to track design team discourse in real time. While LDA has been applied on smaller discourse such as tweets and micro-tweets, recent developments and work in word embeddings, hierarchical topic models, and dynamic topic models have emerged for these purposes [62–64]. LDA was chosen for this specific work for its well-developed and wide utilization of texts of many sizes, accessibility, and previous application on discourse data. The comparison of LDA with these additional topic modeling algorithms lies outside the direct scope of this work (the main goal is not to find the most efficient topic modeling approach), but future work can consider the sensitivity of the corpus size and resulting topics with these other methods. Additionally, the domain, context, and process manager interventions for this problem are quite specific. The problem statement asks participants to design a peanut sheller, and over half of the process manager interventions are tailored toward this specific goal with the design components and design keywords. To fully understand the generalizability of this framework, analysis on different types of problem types and domains can be studied, as well as expanding beyond conceptual design problem-solving. Finally, while this work identifies intended topics shifts in the behavior of the managed teams, additional experimental conditions need to be run to completely isolate this effect from all potential confounding variable, directly link this to overall more effective team performance, and explore other modes of process manager strategies.

6 Conclusion

This work utilizes a topic modeling approach to study the effects of process manager interventions via analysis of design team discourse. The transcripts, collected from a prior research study by the authors, contain discourse of design teams solving a design problem under either the guidance or absence of a human process manager. The inspiration of this topic modeling perspective derives from post-study interviews conducted with the process managers. The goals of imbuing functional concepts into the discourse, and shifting to more relevant topics, serve as some of the primary motivations of the managers for intervening with the design teams.

The topic modeling framework, in this instance LDA, can be leveraged to predict topic mixtures of the team discourse at different segments during problem-solving. Training over a number of topics, optimization algorithms, and training and test sets, 15 topics emerge as the number of topics based on validation perplexity and pointwise mutual information metrics. After this exhaustive training procedure, and corroborative analyses with LSA, the topic model can now be used to transform different intervals of the transcripts into the topic space. The model outputs a probability distribution of the segments of transcripts over the 15 topics. This output allows a more holistic perspective on team discourse, as opposed to mapping to a single topic or coding scheme.

In order to uncover the influence of the process managers, this work covers two analyses to detect the impacts of their interventions. First, the topic model framework is leveraged to predict the overall topic structures between managed and unmanaged teams. This includes both a holistic perspective, through the transformation of the entire transcripts themselves, and a more dynamic perspective, through the transformation of smaller, five-minute intervals over time. An additional analysis studies the direct impacts of the interventions by predicting the topic mixtures immediately prior to and immediately following the interventions. All of these analyses corroborate similar findings and show convergent effects on team discourse, and thus direct impacts of these inventions on design team cognition.

Design team interactions and discourse provide valuable insight into the state and cognition of designers and effectively analyzing them can facilitate the design process. This research provides a computational perspective on not only studying design team verbalizations, but also leveraging communication to detect the effects of design interventions via shifts in topic mixtures. Overall, this work

contributes towards the goal of an automated approach to track design team discourse in real time. Particularly as the collaboration of human and artificial intelligence designers to solve problems becomes more prevalent in practice, being able to computationally track the design state will be critically important to understand what types of interventions may be needed to maximize performance.

Acknowledgment

This work was supported by the Air Force Office of Scientific Research (AFOSR) under Grant No. FA9550-18-0088. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsors. A previous version of this paper was published in the proceedings of the 2020 ASME IDETC Design Theory and Methodology Conference.

Conflict of Interest

There are no conflicts of interest.

Data Availability Statement

The data sets generated and supporting the findings of this article are obtainable from the corresponding author upon reasonable request.

References

[1] Ball, L. J., and Christensen, B. T., 2019, "Advancing an Understanding of Design Cognition and Design Metacognition: Progress and Prospects," Des. Stud., **65**(1), pp. 35–59.
[2] Cross, N., 2001, "Design Cognition: Results From Protocol and Other Empirical Studies of Design Activity," *Design Knowing and Learning: Cognition in Design Education*, C. Eastman, W. Newstatter, and M. McCracken, ed., Elsevier, Oxford, UK, pp. 79–103.
[3] Den Otter, A., and Emmitt, S., 2007, "Exploring Effectiveness of Team Communication: Balancing Synchronous and Asynchronous Communication in Design Teams," Eng. Constr. Archit. Manage., **14**(5), pp. 408–419.
[4] Lloyd, P., Lawson, B., and Scott, P., 1995, "Can Concurrent Verbalization Reveal Design Cognition?," Des. Stud., **16**(2), pp. 237–259.
[5] Stempfle, J., and Badke-Schaub, P., 2002, "Thinking in Design Teams—An Analysis of Team Communication," Des. Stud., **23**(5), pp. 473–496.
[6] Martin, M. J., and Foltz, P. W., 2004, "Automated Team Discourse Annotation and Performance Prediction Using LSA," Proceedings of Human Language Technology and North American Association for Computational Linguistics Conference 2004: Short Papers. Association for Computational Linguistics, Boston, MA, May 2–7, Association for Computational Linguistics, pp. 97–100.
[7] Dong, A., Hill, A. W., and Agogino, A. M., 2004, "A Document Analysis Method for Characterizing Design Team Performance," ASME J. Mech. Des., **126**(3), pp. 378–385.
[8] Dong, A., 2005, "The Latent Semantic Approach to Studying Design Team Communication," Des. Stud., **26**(5), pp. 445–461.
[9] Fu, K., Cagan, J., and Kotovsky, K., 2010, "Design Team Convergence: The Influence of Example Solution Quality," ASME J. Mech. Des., **132**(11), p. 111005.
[10] Dong, A., Kleinsmann, M. S., and Deken, F., 2013, "Investigating Design Cognition in the Construction and Enactment of Team Mental Models," Des. Stud., **34**(1), pp. 1–33.
[11] Linsey, J. S., Markman, A. B., and Wood, K. L., 2012, "Design by Analogy: A Study of the WordTree Method for Problem Re-representation," ASME J. Mech. Des., **134**(4), p. 041009.
[12] Agogué, M., Kazakçi, A., Hatchuel, A., Le Masson, P., Weil, B., Poirel, N., and Cassotti, M., 2014, "The Impact of Type of Examples on Originality: Explaining Fixation and Stimulation Effects," J. Creat. Behav., **48**(1), pp. 1–12.
[13] Song, B., Srinivasan, V., and Luo, J., 2017, "Patent Stimuli Search and Its Influence on Ideation Outcomes," Des. Sci., **3**(1), p. e25.
[14] Fu, K., Chan, J., Cagan, J., Kotovsky, K., Schunn, C., and Wood, K., 2013, "The Meaning of 'Near' and 'Far': The Impact of Structuring Design Databases and the Effect of Distance of Analogy on Design Output," ASME J. Mech. Des., **135**(2), p. 021007.
[15] Fu, K., Cagan, J., Kotovsky, K., and Wood, K., 2013, "Discovering Structure in Design Databases Through Functional and Surface Based Mapping," ASME J. Mech. Des., **135**(3), p. 031006.
[16] Goucher-Lambert, K., Gyory, J. T., Kotovsky, K., and Cagan, J., 2020, "Adaptive Inspirational Design Stimuli: Using Design Output to Computationally Search for Stimuli That Impact Concept Generation," ASME J. Mech. Des., **142**(9), p. 091401.
[17] Gyory, J. T., Cagan, J., and Kotovsky, K., 2019, "Are You Better off Alone? Mitigating the Underperformance of Engineering Teams During Conceptual Design Through Adaptive Process Management," Res. Eng. Des., **30**(1), pp. 85–102.
[18] Gyory, J. T., Cagan, J., and Kotovsky, K., 2018, "Should Teams Collaborate During Conceptual Engineering Design? An Experimental Study," International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Quebec City, Canada, Aug. 26–29, American Society of Mechanical Engineers, Vol. 51845.
[19] Mohr, J. W., and Bogdanov, P., 2013, "Introduction-Topic Models: What They Are and Why They Matter," Poetics, **41**(6), pp. 545–569.
[20] Blei, D. M., 2012, "Probabilistic Topic Models," Commun. ACM, **55**(4), pp. 77–84.
[21] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P., 2004, "The Author-Topic Model for Authors and Documents," Conference on Uncertainty in Artificial Intelligence, Banff, Canada, July 7–11, pp. 487–494.
[22] Blei, D. M., and Lafferty, J. D., 2007, "A Correlated Topic Model of Science," Ann. Appl. Stat., **1**(1), pp. 17–35.
[23] Ball, Z., and Lewis, K., 2020, "Predicting Design Performance Utilizing Automated Topic Discovery," ASME. J. Mech. Des., **142**(12), p. 121703.
[24] Ahmed, F., Fuge, M., and Gorbunov, L. D., 2016, "Discovering Diverse, High Quality Design Ideas From a Large Corpus," International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Charlotte, NC, Aug. 21–24, American Society of Mechanical Engineers, Vol. 50190.
[25] Ahmed, F., and Fuge, M., 2018, "Creative Exploration Using Topic-Based Bisociative Networks," Des. Sci., **4**(12), pp. 1–30.
[26] Bhowmik, T., Niu, N., Savolainen, J., and Mahmoud, A., 2015, "Leveraging Topic Modeling and Part-of-Speech Tagging to Support Combinational Creativity in Requirements Engineering," Requir. Eng., **20**(3), pp. 253–280.
[27] Joung, J., and Kim, H. M., 2021, "Automated Keyword Filtering in Latent Dirichlet Allocation for Identifying Product Attributes From Online Reviews," ASME J. Mech. Des., **143**(8), p. 084501.
[28] Suryadi, D., and Kim, H. M., 2019, "A Data-Driven Approach to Product Usage Context Identification From Online Customer Reviews," ASME J. Mech. Des., **141**(12), p. 121004.
[29] Nijstad, B. A., and Stroebe, W., 2006, "How the Group Affects the Mind: A Cognitive Model of Idea Generation in Groups," Personal. Soc. Psychol. Rev., **10**(3), pp. 186–213.
[30] Hey, J., Linsey, J., Agogino, A. M., and Wood, K. L., 2008, "Analogies and Metaphors in Creative Design," Int. J. Eng. Educ., **24**(2), pp. 283–294.
[31] Yilmaz, S., and Seifert, C., 2010, "Cognitive Heuristics in Design Ideation," Proceedings of 11th International Design Conference, Dubrovnik, Croatia, May 17–20, pp. 1–11.
[32] Isaksen, S. G., 2013, "Facilitating Creative Problem-Solving Groups," http://www.cpsb.com/research/articles/creative-problem-solving/Facilitating-CPS-Groups.html.
[33] Gyory, J. T., Cagan, J., and Kotovsky, K., 2018, "An Exploration of the Effects of Managerial Intervention on Engineering Design Team Performance," International Conference on-Design Computing and Cognition, Como, Italy, July 2–4.
[34] Blei, D. M., Ng, A. Y., and Jordan, M. I., 2003, "Latent Dirichlet Allocation," J. Mach. Learn. Res., **3**(1), pp. 993–1022.
[35] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R., 1990, "Indexing by Latent Semantic Analysis," J. Am. Soc. Inf. Sci., **41**(6), pp. 391–407.
[36] Hofmann, T., 2017, "Probabilistic Latent Semantic Indexing," ACM SIGIR Forum, **51**(2), pp. 211–218.
[37] Ramos, J., 2003, "Using TF-IDF to Determine Word Relevance in Document Queries," Proceedings of the First Instructional Conference on Machine Learning, **242**(1), pp. 29–48.
[38] Kim, T. Y., Min, M., Yoon, T., and Lee, J. H., 2010, "Semantic Analysis of Twitter contents using PLSA, and LDA," SCIS & ISIS, Okayama, Japan, Dec. 8–12.
[39] Rosa, K. D., Shah, R., Lin, B., Gershman, A., and Frederking, R., 2011, "Topical Clustering of Tweets," Proceedings of the ACM SIGIR: SWSM, Beijing, China, July 28.
[40] Mehrotra, R., Sanner, S., Buntine, W., and Xie, L., 2013, "Improving lda Topic Models for Microblogs via Tweet Pooling and Automatic Labeling," Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, July 28–Aug. 1.
[41] Nguyen, V.-A., Boyd-Graber, J., and Resnik, P., 2012, "SITS: A Hierarchical Nonparametric Model Using Speaker Identity for Topic Segmentation in Multiparty Conversations," Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers – Vol. 1, Jeju Island, South Korea, July 8–14, pp. 78–87.
[42] Purver, M., Kording, K., Griffiths, T. L., and Tenenbaum, J. B., 2006, "Unsupervised topic modelling for multi-party spoken discourse," Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, July 17–20, pp. 17–24.
[43] Crain, S. P., Zhou, K., Yang, S. H., and Zha, H., 2012, "Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond," *Mining Text Data*, Springer, Boston, MA.
[44] Handler, A., Denny, M., Wallach, H., and O'Connor, B., 2016, "Bag of What? Simple Noun Phrase Extraction for Text Analysis," Proceedings of 2016

EMNLP Workshop on Natural Language Processing and Computational Social Science, Austin, TX, Nov. 5, pp. 114–124.
[45] Loper, E., and Bird, S., 2004, "NLTK: The Natural Language Toolkit," Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions, Association for Computational Linguistics, Barcelona, Spain, July 21–26.
[46] Huang, L., Ma, J., and Chen, C., 2017, "Topic Detection From Microblogs Using T-LDA and Perplexity," Proceedings—2017 24th Asia-Pacific Software Engineering Conference Workshops (APSECW), Nanjing, China, Dec. 4–8, pp. 71–77.
[47] AlSumait, L., Barbará, D., and Domeniconi, C., 2008, "On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking," Proceedings—IEEE International Conference on Data Mining, ICDM, Pisa, Italy, Dec. 15–19, IEEE, pp. 3–12.
[48] Heinrich, G., 2005, "Parameter Estimation for Text Analysis," Technical Report, **1**(1), pp. 1–36.
[49] Merriam-Webster, M. W. S. C., 1977, *Dictionary*, G & C Merriam Company, 830., Springfield, MA.
[50] Kullback, S., and Leibler, R. A., 2010, "On Information and Sufficiency," Ann. Math. Stat., **22**(1), pp. 79–86.
[51] Pérez-Cruz, F., 2008, "Kullback-Leibler Divergence Estimation of Continuous Distributions," IEEE International Symposium on Information Theory—Proceedings, Toronto, ON, Canada, July 6–11, pp. 1666–1670.
[52] Shlens, J., 2014, "Notes on Kullback-Leibler Divergence and Likelihood." arXiv preprint arXiv:1404.2000.
[53] Kapoor, R., Gupta, R., Son, L. H., Jha, S., and Kumar, R., 2018, "Boosting Performance of Power Quality Event Identification With KL Divergence Measure and Standard Deviation," Measurement, **126**(1), pp. 134–142.
[54] Foulds, J., Boyles, L., DuBois, C., Smyth, P., and Welling, M., 2013, "Stochastic Collapsed Variational Bayesian Inference for Latent Dirichlet Allocation," Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, Aug. 11–14, ACM, pp. 446–454.
[55] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J., 2013, "Stochastic Variational Inference," J. Mach. Learn. Res., **14**(1), pp. 1303–1347.
[56] Griffiths, T. L., and Steyvers, M., 2004, "Finding Scientific Topics," Proc. Natl. Acad. Sci. USA, **101**(1), pp. 5228–5235.
[57] Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W., 2009, "On Smoothing and Inference for Topic Models," Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada, June 18–21, AUAI Press, pp. 27–34.
[58] Teh, Y. W., Newman, D., and Welling, M., 2007, "A collapsed variational bayesian inference algorithm for latent dirichlet allocation," Adv. Neur. Infor. Proc. Sys., **19**(1), pp. 1353–1360.
[59] Airoldi, E. M., Blei, D., Erosheva, E. A., and Fienberg, S. E., 2014, *Handbook of Mixed Membership Models and Their Applications*, CRC Press, Boca Raton, FL.
[60] Newman, D., Baldwin, T., Cavedon, L., Huang, E., Karimi, S., Martinez, D., Scholer, F., and Zobel, J., 2010a, "Visualizing Search Results and Document Collections Using Topic Maps," Web Semantics, **8**(2–3), pp. 169–175.
[61] Newman, D., Lau, J. H., Grieser, K., and Baldwin, T., 2010b, "Automatic Evaluation of Topic Coherence," Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, CA, June 2–4, pp. 100–108.
[62] Blei, D. M., and Lafferty, J. D., 2006, "Dynamic Topic Models," Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, June 25–29.
[63] Pujara, J., and Skomoroch, P., 2012, "Large-Scale Hierarchical Topic Models." NIPS Workshop on Big Learning. Vol. **128**.
[64] Liu, Y., Liu, Z., Chua, T. S., and Sun, M., 2015, "Topical Word Embeddings," Twenty-Ninth NAAI Conference on Artificial Intelligence, Austin, TX, Jan. 25–30.