JOURNAL of MARKETING RESEARCH

A Poisson Factorization Topic Model for the Study of Creative Documents (and their Summaries)

Journal:	*Journal of Marketing Research*
Manuscript ID	JMR.19.0088.R3
Manuscript Type:	Special Issue Revised Submission
Topics and Methods:	Statistics < Theoretical Foundation, Bayesian estimation < Theoretical Foundation, Topic models, Natural language processing, Creativity

A Poisson Factorization Topic Model for the Study of Creative Documents (and their Summaries)

The authors propose a topic model tailored to the study of creative documents (e.g., academic papers, movie scripts). They extend Poisson Factorization in two ways. First, the creativity literature emphasizes the importance of novelty in creative industries. Accordingly, they introduce a set of residual topics that represent the portion of each document that is not explained by a combination of common topics. Second, creative documents are typically accompanied by summaries (e.g., abstracts, synopses). Accordingly, they jointly model the content of creative documents and their summaries, and capture systematic variations in topic intensities between the documents and their summaries. The authors validate and illustrate the model in three domains: marketing academic papers, movie scripts, and TV show closed captions. They illustrate how the joint modeling of documents and summaries provides some insight into how humans summarize creative documents, and enhances our understanding of the significance of each topic. They show that their model produces new measures of distinctiveness which can inform the perennial debate on the relation between novelty and success in creative industries. Finally, they show how the proposed model may form the basis for decision support tools that assist humans in writing summaries of creative documents.

With the digitization of the economy, people are both producing and consuming more and more creative content. On the supply side, according to Florida (2014), more than 40 million of Americans (or approximately one third of the employed population) belong to the "creative class." This class includes people in science and engineering, education, arts, entertainment, whose primary economic function is to create new ideas, new technology, and new creative content. On the demand side, the average american spends approximately 12 hours per day consuming media (Statista, 2017), and the media and entertainment industry alone is valued at approximately $2 trillion globally (Statista, 2018).

In this paper, we use the term "creative document" to refer to any written document that describes the output of a creative process. Examples include academic papers, fiction books, movie and TV show scripts, plays, business models, and new product descriptions. In contrast, non-creative documents include news articles, instructions manuals, etc. In addition to being managerially relevant, creative documents have captured the interest of academics. Several studies have attempted to identify correlates of success in creative industries, and in particular the link between the distinctiveness of a creative document and its success (e.g., the link between the distinctiveness of an academic paper and its number of citations).

Studying creative documents at a large scale in a scientific manner has been historically challenging, due to the unstructured nature of the data contained in these documents. With the development of natural language processing tools such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) or Poisson Factorization (Canny, 2004), it has become possible to systematically extract text-based topics and features from creative documents. Although some papers have applied variations of traditional topic models to the study of creative documents (e.g., Eliashberg et al., 2007, 2014; Berger and Packard, 2018; Toubia et al.,

2018), we argue that these traditional topic models fail to capture at least two essential aspects of creative documents.

First, the creativity literature has shown that *novelty* is a key construct when it comes to creative content. Traditional topic models perform dimensionality reduction by approximating each document using a set of topics, which are common across all documents in the corpus. With a traditional topic model, the distinctiveness of a document may be measured by the distinctiveness of its combination of topics. However, traditional topic models fail to capture another aspect of distinctiveness: the extent to which a document may *not* be explained by common topics. As such, we argue that traditional topic models are limited in their ability to provide rich measures of distinctiveness which may inform the debate on the link between novelty and success in creative industries.

Second, creative documents are often accompanied by *summaries*. For example, academic papers are accompanied by abstracts, books and movies by synopses, new products by short descriptions, business plans by executive summaries, etc. Summaries play a key role in the market, by helping consumers extract information from creative products more efficiently and decide which products to consume. For example, a consumer may be enticed to buy a book or watch a movie based on a synopsis, or to buy a new product based on its short description. One may argue that summaries serve as "lubricant" in the market for creative content, and soften competition by making it easier for consumers to decide which products to consume.[1] Traditional topic models do not capture the relation between a document and its summary. We argue that modeling and quantifying the process by which humans summarize creative documents is not only interesting from an academic perspective, but that it also offers

[1]We thank Anthony Dukes for this insight.

practical benefits. From the perspective of extracting meaningful, interpretable topics from a corpus of creative documents, summaries may be viewed as shorter documents produced by humans who invested time and effort to determine which topics in a creative document are "essential" enough to be included in the document's summary. As such, summaries have the potential to improve our understanding of the significance of each topic. Moreover, modeling the summarization process opens the door for the development of computer-based tools to assist authors and marketers in creative industries in writing summaries of creative documents. For example, by identifying characteristics of summaries that correlate with success in a specific creative industry, we can advise authors to emphasize certain topics in their summaries.

Motivated by these two characteristics of creative documents, we propose a topic model tailored to the study of creative documents. Our contribution in this paper is primarily methodological. Our model extends Poisson Factorization in two ways. First, we account not only for the portion of a document that may be explained by topics that are common across documents, but also the "residual" (or "outside the cone" - see the geometric interpretation) portion that is not explained by combinations of these common topics. Second, we jointly model the content of creative documents and their summaries. The model represents systematic variations in the extent to which each common topic, as well as each "residual" topic, appears in summaries compared to full documents.

While topic models have been applied to creative documents, to the best of our knowledge our model is the first topic model specifically tailored for creative documents. Our model offers at least three additional benefits to academics and practitioners, which are not offered by traditional topic models, and which we illustrate in this paper. First, each topic

estimated by our model comes with a variable that quantifies the extent to which the topic was deemed "summary worthy" by the humans who wrote the summaries of the documents in the corpus. We illustrate how this additional layer of information provides some insight into the process by which humans summarize creative documents in a particular domain, and enhances our understanding of the significance of each topic. Second, for academics and practitioners interested in participating in the ongoing debate on the link between distinctiveness and success of creative products, we show that our model provides various measures of distinctiveness, which have the potential to uncover new insight into correlates of success in creative industries. We explore empirically in our three datasets the relation between three measures of distinctiveness and various success measures (i.e., number of citations of academic papers, movie and TV show ratings, movie return on investment). Third, we show that our model may serve as the basis for interactive decision support tools that assist humans in writing summaries of creative documents. The development of such tools may be informed by an empirical analysis of correlates of success in the target industry. For example, we find that marketing academic papers whose abstracts put relatively more emphasis on the "outside the cone" content in the paper, tend to have more citations. Accordingly, our model can help authors identify the "outside the cone" content in their paper, and emphasize it in their abstract. We develop a proof of concept for such a tool.

Relevant Literatures

The study of creativity in various domains, from scientific discovery (e.g., Uzzi et al., 2013) to linguistics (e.g., Giora, 2003) and innovation (Toubia and Netzer, 2017), has suggested that creativity lies in the optimal balance between novelty and familiarity. For example,

Ward (1995) argues that "truly useful creativity may reflect a balance between novelty and a connection to previous ideas." Furthermore, based on previous research from a wide range of domains (e.g., Mednick, 1962; Finke et al., 1992), Toubia and Netzer (2017) show that when attempting to quantify familiarity and novelty in a document using text analysis, researchers should focus on novel vs. familiar *combinations* of words, rather than words that themselves appear more or less frequently.

These insights inform our modeling approach. We adopt a natural language processing approach, which captures topics defined as combinations of words. Our model nests and extends Poisson Factorization. Previous applications of Poisson Factorization to the study of text documents include Canny (2004) and Gopalan et al. (2014). For example, Gopalan et al. (2014) study how researchers rate academic papers, by modeling documents and researcher preferences as latent vectors in a topic space. The model we propose builds on Gopalan et al. (2014)'s model and differs from it in a few important ways. We model the content of full documents and their summaries, rather than modeling the content of documents and consumers' preferences for these documents. These different objectives give rise to very different data, model specifications and data generating processes. We jointly model the content of documents and their summaries, we explicitly model "residual" topics, and we model how residual topics are represented in summaries; none of which is performed by Gopalan et al. (2014)'s model. We also use offset variables in a novel way, to capture systematic variations in topic intensities in full documents vs. summaries. As noted in the introduction, several papers have used *extant* topic models to study creative documents (e.g., Eliashberg et al., 2007, 2014; Berger and Packard, 2018; Toubia et al., 2018). However, to the best of our knowledge our model is the first topic model tailored to the study of creative

documents.

We note that most applications of topic modeling in the marketing literature have used Latent Dirichlet Allocation (LDA, Blei et al., 2003) or extensions thereof (e.g., Tirunillai and Tellis, 2014; Büschken and Allenby, 2016; Puranam et al., 2017; Liu and Toubia, 2018; Toubia et al., 2018; Zhong and Schweidel, 2018). The basic LDA model shares many similarities with the basic Poisson Factorization model, although previous research has suggested that Poisson Factorization tends to fit data better (Canny, 2004; Gopalan et al., 2013, 2014). Our choice of Poisson Factorization was primarily driven by the attractive conjugacy property of this approach. Indeed, our model remains conditionally conjugate, despite the additional complexities resulting from jointly modeling the content of documents and summaries while explicitly capturing residual content.[2]

Despite the importance of summaries in the commercialization of creative content, summarization has received very little attention in the marketing literature. Nevertheless, text summarization is a substantial subfield of computer science (see, for example, Radev et al., 2002; Nenkova and McKeown, 2012; Allahyari et al., 2017; Yao et al., 2017). However, computer scientists have focused mostly on *automatic* text summarization, where a summary is produced without any human intervention. This is typically done by identifying and selecting a subset of the sentences in the original document, a process called *extractive summarization* (Allahyari et al., 2017). Such text summarization tools are useful to summarize large numbers of documents (e.g., news articles) on a regular basis, quickly and efficiently (McKeown and Radev, 1995; Radev and McKeown, 1998). In contrast, we focus on situations in which

[2]Word embedding (Mikolov et al., 2013, 2017) has recently emerged as another popular natural language processing approach. Word embedding typically does *not* extract topics from text and does not assign topic intensities to documents. Hence it is not directly relevant to our goal of developing a topic model tailored to the study of creative documents. However, by capturing the *context* around each word, word embedding may be better suited for studying the *structure* of creative document, which we leave for future research.

summaries provide additional content written by humans, from which valuable insights might be learned. In terms of practical applications, we envision computers not as a replacement for, but rather as an aid to humans, and consider decision support tools that assist humans in writing summaries of creative documents. Our different perspective on summarization also translates into methodological differences. Some papers have applied topic modeling to text summarization, sometimes introducing document-specific topics that capture unique content in each document, which should be included in the summary (Daumé III and Marcu, 2006; Haghighi and Vanderwende, 2009; Delort and Alfonseca, 2012). These document-specific topics are similar in spirit to the residual topics in our model. However, given their focus on extractive summarization, unlike our model, these models did not consider summaries as an additional source of information, they did not model the content of summaries, and they did not include summaries in their training data.

Proposed Model

Model Foundation: Poisson Factorization

We index creative documents by $d = 1, ...D$, and words in the vocabulary by $v = 1, ...L$. We denote as w_{dv} the number of times word v appears in document d. In *standard* Poisson Factorization (Canny, 2004; Gopalan et al., 2014), the assumed data generating process would be as follows:

1. For each regular topic $k = 1, ...K$:
    - For each word v, draw $\beta_{kv} \sim Gamma(\alpha_1, \alpha_2)$
2. For each document $d = 1, ...D$:

- For each topic, draw topic intensity $\theta_{dk} \sim Gamma(\alpha_3, \alpha_4)$
- For each word v, draw word count $w_{dv} \sim Poisson(\sum_k \theta_{dk}\beta_{kv})$

In order to gain intuition for this base model, recall that the sum of independent Poisson-distributed random variables is a Poisson variable. Hence, according to Poisson Factorization, the number of occurrences of word v in document d, $w_{dv} \sim Poisson(\sum_k \theta_{dk}\beta_{kv})$, may be thought of as the sum of K independent Poisson variables (called auxiliary variables, e.g., Gopalan et al., 2014): $z_{dv,k} \sim Poisson(\theta_{dk}\beta_{kv})$. These variables capture the number of occurrences of word v in document d associated with each topic k, such that $w_{dv} = \sum_k z_{dv,k}$. The distribution of each auxiliary variable $z_{dv,k}$ is influenced by the product of two terms: θ_{dk} represents the intensity of topic k in document d; β_{kv} represents the weight of word v in topic k.

One can also interpret Poisson Factorization geometrically. (To the best of our knowledge, the following geometric interpretation of Poisson Factorization is new to the literature.) Topics and documents may be represented in the Euclidean space defined by the words in the vocabulary. That is, topic k may be represented by a $L \times 1$ vector $\beta_k = \{\beta_{kv}\}_v$ that captures the weights on each word in the topic. Similarly, document d may be represented by a $L \times 1$ vector $w_d = \{w_{dv}\}_v$ that contains the number of occurrences of each word observed in the document. According to Poisson Factorization, the expected value of this vector is given as: $E(w_d) = \sum_k \theta_{dk}\beta_k$. (Recall that the expected value of a variable with a Poisson distribution is the rate of the distribution.) That is, the expected number of occurrences of words in the document may be written as a positive combination of the vectors $\{\beta_k\}_k$ that represent topics in the word space, where the weights are the topic intensities $\{\theta_{dk}\}_k$. In this illustration, for simplicity we focus on expected values.

Mathematically, the positive combinations of the set of topic vectors, $\{\sum_k \theta_{dk}\beta_k,\ \theta_{dk} \geq 0\}$, form a cone in the Euclidean space defined by the words in the vocabulary. This means that Poisson Factorization may be viewed as approximating each document by projecting it onto the cone defined by the topic vectors. We provide an illustration of this geometric interpretation in Figure 1. In this figure, for illustration purposes we consider a vocabulary that consists of three words, and we assume three topics (in practice the number of topics should be much smaller than the number of words in the vocabulary). This figure illustrates the cone defined by positive combinations of the three topics. It also shows the example of one document represented by a vector in the same space, and how Poisson Factorization projects this document vector onto the cone defined by the topics.[3] (Note that in reality the projection is not orthogonal, due to the prior on the parameters.)

<INSERT FIGURE 1 ABOUT HERE>

In sum, the primary focus of traditional topic models such as Poisson Factorization is to understand topics that are common across documents in a corpus, and to quantify the intensity with which each topic is featured in each document. In doing so, Poisson Factorization approximates each document as a positive combination of common topics.

Residual Topics

Our model extends Poisson Factorization in two ways. First, we capture "outside the cone" content, by introducing one "residual topic" associated with each document. For each document d, we introduce a topic β_d^{res} that is unique to this document. The weight of this

[3]We note that in the case of LDA, documents are approximated by convex combinations of the topics. Hence, cones would be replaced with simplexes in this geometric interpretation.

topic on each word v is assumed to have a Gamma prior, similar to the other, "regular" topics: $\beta_{dv}^{res} \sim Gamma(\alpha_1, \alpha_2)$. We model the number of occurrences of word v in document d as: $w_{dv} \sim Poisson(\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} + \beta_{dv}^{res})$, where the superscript reg refers to the regular topics (β_{kv}^{reg} is common across all documents in the corpus).[4] The residual topic represents the residual content in document d. To the best of our knowledge, this paper is the first to introduce residual topics in Poisson Factorization.

The introduction of this residual topic was motivated by the creativity literature, in an attempt to account for distinct content in the document. One may wonder whether the residual topic is simply "noise." In a later section, we empirically test whether the residual topic indeed relates to the success of creative documents in ways that are predicted by the creativity literature. If this topic were "just noise," we should find no systematic relation with the success of creative documents. Theoretically, we note that the model still includes "noise," above and beyond the residual topics. Indeed, the number of occurrences of each word remains stochastic and governed by a Poisson distribution. We also note that the prior induces sparsity and trades off fit with the complexity of the model. As a result, the expected value of the number of occurrences of each word according to the model does not perfectly fit the observed value, even in the presence of residual topics.

Figure 1 illustrates geometrically how the vector corresponding to a document is decomposed into two vectors: the "inside the cone" component that projects the document vector onto the cone defined by the regular topics, and the "outside the cone" component that closes the gap between the original vector and the projection. (Again, this simple illustration focuses on expected values and ignores the effect of the prior - our actual model produces a

[4]Note that the intensity of the residual topic θ_d^{res} is implicitly set to 1, for identification purposes.

distribution of word occurrences and fit is not perfect due to the sparsity-inducing prior.)

Offset Variables

The second way in which we extend Poisson Factorization is that we jointly model the content of creative documents and their summaries. To that end, we introduce a set of "offset" variables, that capture how topics are weighed in summaries, compared to full documents. The topic intensities in the summary of a creative document may not be the same as the topic intensities in the full document. First, some regular topics may be typically judged by the authors of summaries as being more or less worthy of being featured in a document's summary. This should translate into systematic differences across regular topics in the way they are weighed in summaries vs. full documents. For example, topics that relate to data analysis (respectively, substantive findings) may be relatively under-weighed (respectively, over-weighed) in the abstracts of academic papers vs. the full papers. In order to capture and quantify such phenomenon, we allow each regular topic k to have its own "offset" variable, ϵ_k^{reg}. Second, "inside the cone" and "outside the cone" content may be weighed differently in summaries vs. full documents. Accordingly, we also introduce an offset variable for each residual topic, ϵ_d^{res}. More precisely, we model the number of occurrences of word v in the summary of document d as: $w_{dv}^{summary} \sim Poisson(\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} \epsilon_k^{reg} + \beta_{dv}^{res} \epsilon_d^{res})$. That is, the topic intensity of regular topic k in the full document, θ_{dk}^{reg}, is multiplied by the offset variable ϵ_k^{reg} in the summary. Similarly, the intensity of the residual topic is multiplied by the offset variable ϵ_d^{res}. By specifying Gamma priors on the offset variables, we preserve the conditional conjugacy of the model, i.e., the posterior distribution of each variable conditional on the other variables and the data, is given in closed form.

A perennial issue with traditional topics models is the difficulty of interpreting topics, resulting from the unsupervised nature of these models. Offset variables provide an additional layer of information that helps understand the significance of each topic, by giving it a "score" that captures the extent to which humans decide to include this topic when writing summaries of creative documents in the domain under study. While offset variables have been used for different purposes in previous applications of Poisson Factorization (e.g., Gopalan et al., 2014), to the best of our knowledge this paper, being the first to use Poisson Factorization to jointly model documents and their summaries, is also the first to use offset variables to capture how the intensities of topics vary between documents and summaries. In Web Appendix F, we further explore the impact of introducing offset variables, by estimating an alternative version of the model that does not include these variables. We find that the topics learned by this alternative model are substantively different from the topics learned by the proposed model. In the proposed model, topics are defined as groups of words that not only tend to appear together, but that that also tend to appear with the same relative frequency in summaries vs. full documents. Accordingly, the presence of offset variables affects the topics learned from the model.

Data Generating Process

Putting all these pieces together, the data generating process for our model is as follows:

1. For each regular topic $k = 1, ...K$:

   - For each word v, draw $\beta_{kv}^{reg} \sim Gamma(\alpha_1, \alpha_2)$
   - Draw offset variable $\epsilon_k^{reg} \sim Gamma(\alpha_5, \alpha_6)$

2. For each residual topic $d = 1, ...D$:

   - For each word v, draw $\beta_{dv}^{res} \sim Gamma(\alpha_1, \alpha_2)$
   - Draw offset variable $\epsilon_d^{res} \sim Gamma(\alpha_5, \alpha_6)$

3. For each document $d = 1, ...D$:

   - For each regular topic, draw topic intensity $\theta_{dk}^{reg} \sim Gamma(\alpha_3, \alpha_4)$
   - For each word v, draw word count $w_{dv} \sim Poisson(\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} + \beta_{dv}^{res})$

4. For each document summary $d = 1, ...D$:

   - For each word v, draw word count $w_{dv}^{summary} \sim Poisson(\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} \epsilon_k^{reg} + \beta_{dv}^{res} \epsilon_d^{res})$

Estimation Using Variational Inference

In order to estimate the model, we start by defining auxiliary variables that allocate the occurrences of each word v in each document d across the various topics: $z_{dv,k}^{reg} \sim Poisson(\theta_{dk}^{reg} \beta_{kv}^{reg})$; $z_{dv}^{res} \sim Poisson(\beta_{dv}^{res})$, such that $w_{dv} = \sum_k z_{dv,k}^{reg} + z_{dv}^{res}$. Similar variables are defined for the summaries: $z_{dv,k}^{sum,reg} \sim Poisson(\theta_{dk}^{reg} \beta_{kv}^{reg} \epsilon_k^{reg})$; $z_{dv}^{sum,res} \sim Poisson(\beta_{dv}^{res} \epsilon_d^{res})$, such that $w_{dv}^{summary} = \sum_k z_{dv,k}^{sum,reg} + z_{dv}^{sum,res}$. With the addition of these auxiliary variables, the model has the attractive property of being conditionally conjugate, i.e., the posterior distribution of each parameter conditional on the other parameters and the data, is given in closed form. The model could be estimated using Gibbs sampling. Instead, in order to speed up computations and improve scalability, we estimate it using Variational Inference (Blei et al., 2016). Details are provided in Web Appendix B.

Selecting the Number of Topics

The number of topics could be selected using cross-validation, to achieve minimum perplexity. Instead, we use a simpler approach advocated by Gopalan et al. (2014). That is, we set the number of topics K to a large number (like these authors, we use $K = 100$), with the realization that some of these topics will be "flat," i.e., such that all topic weights β_{kv}^{reg} are very small and similar across words and all topic intensities θ_{dk}^{reg} are very small and similar across documents. We set the same value of $K = 100$ for all benchmarks. These "flat" topics emerge as a result of the Gamma priors on topic weights and topic intensities, which induce sparsity. In other words, the model automatically attempts to explain the data with few topics, and corrects for values of K that are larger than needed. This means that the number of non-flat topics is influenced by the prior parameters $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6\}$. In this paper, we follow Gopalan et al. (2014) and set $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4\alpha_5 = \alpha_6 = 0.3$. In Web Appendix E, we test a more/less diffuse prior and report how the number of non-flat topics (as well as the distinctiveness measures introduced later) vary in each dataset when the prior is changed.

Extension: Dynamic Topics

In Web Appendix G, we introduce a dynamic extension of this model. We introduce dynamics in a manner inspired by Blei and Lafferty (2006). We model each topic as having a base version, and we introduce a set of time-specific offset variables that capture the evolution of each topic over discrete time periods. In each time period, the weights of each topic are assumed to be equal to the weights in the previous period, plus a set of offset variables specific to that topic and that time period. This extension is also estimated

using variational inference. We apply it to our marketing academic papers datasets, which contains all papers published in a set of journals over 6 years. We find that the introduction of dynamics does not change the conclusions from our empirical analysis.

Empirical Applications

Datasets

We apply our model to three datasets. In each dataset, all documents were preprocessed following standard steps in natural language processing: eliminate non-English characters and words, numbers, and punctuation; tokenize the text (i.e., break each document into individual words or tokens); remove common stop words; remove tokens (words) that contain only one character. No stemming or Lemmatization was performed. In each dataset, we randomly split the set of documents into two samples: a calibration set with 75% of the documents and a validation set with 25% of the documents.

We construct the vocabulary of words in each dataset based on the *full documents* in the *calibration* set only (i.e., the summaries and the validation documents are not used to select the vocabulary). We compute the term frequency (tf) for each word, i.e., the total number of occurrences of the word across all training documents. We remove words that appear fewer than 100 times across documents (for movies, given the smaller sample size, we use a cutoff of $tf < 65$). Next, we compute the tf-idf of each word w, defined as: $tf\text{-}idf(w) = tf(w) \times log(\frac{N}{df(w)})$, where $df(w)$ is the document frequency for word w, defined as the number of documents in which word w appears at least once. The final vocabulary consists of the 1,000 words with the highest tf-idf, i.e., we remove words that appear too frequently and words that appear too infrequently (Blei and Lafferty, 2009). In Web Appendix H, we

run all models with vocabularies of 500 words and with vocabularies of 2,000 words (still selecting words based on *tf-idf*). We find that results are qualitatively similar to the ones obtained with 1,000 words, although changing the vocabulary size does change the estimated measures of distinctiveness introduced later. A simulation study, reported in Web Appendix H, confirms that distinctiveness measures should indeed be affected by vocabulary size.

Our first dataset consists of the full texts (excluding the abstracts, bibliographies and appendices) and the abstracts of all 1,333 research papers published in *Journal of Consumer Research*, *Journal of Marketing*, *Journal of Marketing Research*, and *Marketing Science*, between 2010 and 2015. Most of the papers were downloaded in PDF format. Some spelling errors occurred while converting PDF files to text files, hence, a spelling corrector was trained based on the autocorrection package in Python and applied before preprocessing the data. Table 1 reports descriptive statistics for all datasets, after preprocessing.

For our second dataset, we collect the scripts and synopses of 858 movies released in the US, for which scripts were available on the internet movie script database (imsdb.com) and synopses were available on the internet movie database (imdb.com). Words corresponding to names of locations, persons and organizations were identified using the Stanford Named Entity Recognition classifier, and removed from the data before preprocessing.

For our third dataset, we collaborated with a major global media company, who was interested in creating a "knowledge graph" for their extensive library of TV content, i.e., identifying a set of meaningful, interpretable topics that describe each TV show episode in order to classify its content. The company made available to us the collection of closed captions for 26,561 unique TV show episodes, which constitute most of the company's catalog of US-based, English-language TV show episodes. The company decided to work with closed

captions because they are available systematically and consistently for all episodes, as they are required by the FCC. The company also made available to us the synopses of all TV show episodes, which are part of its internal programming system. As in the previous dataset, words corresponding to names of locations, persons and organizations were removed from the data before preprocessing.

<INSERT TABLE 1 ABOUT HERE>

Fit and Predictive Performance

Benchmarks

The proposed model extends Poisson Factorization in two ways. First, we model "residual" topics, which are unique to each document. Second, we allow the topic intensities in summaries to differ from the topic intensities in main documents. In order to test the benefits of these two extensions, we test a series of nested models. All benchmarks are estimated using variational inference, with the same convergence criterion and hyperparameters.

The first benchmark we consider is a nested model that does not include residual topics. This benchmark is a nested version of the proposed model, in which $\{\beta_d^{res}, \epsilon_d^{res}\}_d$ are constrained to be 0. This benchmark still includes offset variables for the regular topics $\{\epsilon_k^{reg}\}_k$. It allows us to explore the benefit of modeling "outside the cone" content using residual topics. The second benchmark includes residual topics, but constrains all offset variables, $\{\epsilon_k^{reg}\}_k$ and $\{\epsilon_d^{res}\}$, to be equal to each other. That is, this benchmark assumes that the relative intensities of topics in summaries are the same as in the main documents. This benchmark is implemented by replacing the offset variables with a single variable ϵ. This benchmark allows us to explore the benefit of allowing the relative topic intensities in

summaries to differ from those in the main documents. The third nested benchmark does not include residual topics ($\{\beta_d^{res}, \epsilon_d^{res}\}_d$ are set to 0) and constrains all offset variables on the regular topics $\{\epsilon_k^{reg}\}_k$ to be equal. That is, this benchmark is similar to a basic Poisson Factorization model that would assume that documents and their summaries have the same relative topic intensities. The fourth and final nested benchmark does not contain any regular topic, but only residual topics. That is, this benchmark does not attempt to learn topics that are shared across all documents, but rather treats each document as completely unique and learns one residual topic for each document. This benchmark is a special case of the proposed model, in which the number of regular topics K is set to 0.

Finally, we consider LDA as a non-nested benchmark, due to its popularity. Because LDA does not include offset variables, the topic intensities in the summary of a document are assumed to be the same as in the full document. In addition, LDA does not include residual topics. Details of the LDA benchmark are provided in Web Appendix D.

Measures of Fit

We estimate each model on the full texts and summaries of the calibration documents in each dataset. The output from our model and any of its nested benchmark may be summarized by computing a vector of fitted Poisson rates $\tilde{\lambda}_d = \{\tilde{\lambda}_{dv}\}_v$ for each document, which govern the number of occurrences of each word in the document:

$$\tilde{\lambda}_d = \sum_k \theta_{dk}^{reg} \beta_k^{reg} + \beta_d^{res} \quad (1)$$

In addition, for each document, we can construct fitted Poisson rates for the number of occurrences of words in the document's summary:

$$\tilde{\lambda}_d^{summary} = \sum_k \theta_{dk}^{reg} \beta_k^{reg} \epsilon_k^{reg} + \beta_d^{res} \epsilon_d^{res} \tag{2}$$

In order to compare our model to LDA, we transform these Poisson rates into multinomial distributions $\tilde{\phi}_d$ and $\tilde{\phi}_d^{summary}$, where $\tilde{\phi}_{dv} = \frac{\tilde{\lambda}_{dv}}{\Sigma_{v'} \tilde{\lambda}_{dv'}}$ captures the probability that a given word in the document is equal to word v, and similarly for $\tilde{\phi}_d^{summary}$.

We measure fit using the standard measure of perplexity (Blei et al., 2003). Given a set of full documents D_{test} with a total of N words, where the word distribution of each document d is fitted by the $1 \times L$ vector $\tilde{\phi}_d$ and where $\{obs\}$ represent the indices of the words observed in the documents, the perplexity score is given as follows:

$$Perplexity = exp(-\frac{\sum_{d \in D_{test}} \sum_{obs \in d} log(\tilde{\phi}_{d,obs})}{N}) \tag{3}$$

Perplexity is defined similarly for the document summaries:

$$Perplexity^{summary} = exp(-\frac{\sum_{d^{summary} \in D_{test}} \sum_{obs \in d^{summary}} log(\tilde{\phi}_{d,obs}^{summary})}{N^{summary}}) \tag{4}$$

where $N^{summary}$ is the total numbers of words in the summaries and $obs \in d^{summary}$ refers to the words observed in the summary of document d. Note that perplexity is equivalent to the inverse of the geometric mean of the per-word likelihood. Lower scores indicate better fit.

For each model we also estimate the intensities on regular topics $\{\theta_{d^{val}k}^{reg}\}_k$ and the residual topic weights $\{\beta_{d^{val}v}^{res}\}_v$ for each validation document d^{val}, based on the text of this document and the parameters estimated from the calibration sample. Details are provided in Web

Appendix C. Following Equation 3, we compute a perplexity score for the full texts of the validation documents.

Therefore, our in-sample fit measures consist of the perplexity scores for the full texts of the calibration documents, the summaries of the calibration documents, and the full texts of the validation documents. In addition, we report in Web Appendix H the DIC for each benchmark, and find that it is lowest for the full model, in all three datasets.

Measure of Predictive Performance

The predictive task we consider is that of predicting the content of the summary of a validation document, given the full text of this document and the model parameters estimated on the set of calibration documents. Consider a validation document d^{val} for which we estimate the intensities on the regular topics and the residual topic weights (as described above and detailed in Web Appendix C), and for which we attempt to predict the content of the summary. Poisson rates for the summary of document d^{val} are predicted according to Equation 2.[5] These rates capture the occurrences of words in the summary, predicted based on the full text of the document, given the model. Following Equation 4, we compute a perplexity score for the summaries of the validation documents. This is our measure of predictive performance.

Results

We report the performance of the proposed model, the nested benchmarks, and LDA on each of our three datasets, in Table 2. The comparisons between benchmarks are similar

[5]We use the average ϵ_d^{res} from the validation documents in Equation 2 instead of $\epsilon_{d^{val}}^{res}$ when predicting summaries of out-of-sample documents based on their full texts, as the estimation of $\epsilon_{d^{val}}^{res}$ would require access to very summary we are trying to predict.

across datasets. We see that the proposed model performs best in terms of fitting the summaries of calibration documents and predicting the summaries of validation documents. We also see that the "No residual topic" benchmark usually performs worse than the "ϵ constant" benchmark. This suggests that the better performance of the full model is driven primarily by the inclusion of "residual" topics rather than by allowing various topics to be weighed differently in summaries compared to the full texts. One exception is the TV shows dataset, in which the "No residual topic" benchmark performs better than "ϵ constant" at predicting the content of the summaries of validation documents. As shown later, this dataset is the one that features the highest variation in the offset variables $\{\epsilon_k^{reg}\}$ across regular topics. It is therefore not surprising that assuming that ϵ is constant is more detrimental in that dataset.

The "Residual topics only" benchmark, not surprisingly, performs best in terms of fitting the full documents. This benchmark does not attempt to learn any topic across documents, i.e., it does not generate any substantive insight. In addition, the fit on the full documents comes at the expense of fitting or predicting the content of the summaries of documents. Interestingly, this benchmark performs similarly to the "ϵ constant" benchmark at predicting the content of validation summaries. Both of these benchmarks include residual topics, and they both ignore differences across topics in their propensity to be featured in summaries vs. full documents. This is particularly detrimental in the TV shows datasets, in which offset variables vary the most across topics.

Finally, we see that LDA performs very similarly to the benchmark that has no residual topic and constant offset variables. This benchmark is equivalent to traditional Poisson Factorization, which has many similarities with LDA.

<INSERT TABLE 2 ABOUT HERE>

In Web Appendix H, we test an alternative measure of predictive performance, in which we randomly holdout a subset of the word occurrences in each validation document, which are predicted based on the parameter estimates and the other words in the document. In this scenario, the content of validation summaries is predicted based only on a subset of the words in the full document. We find that the full model performs best in terms of predicting the heldout portion of validation documents and the summaries of validation documents, with the exception of the marketing academic papers dataset in which the "Residual topics only" benchmark performs slightly better at predicting the heldout portion of validation documents.

In sum, these results suggest it is reasonable to extend Poisson Factorization to study creative documents and their summaries, by capturing residual content and capturing systematic differences in topic intensities in summaries vs. full documents using offset variables. In the following three sections, we illustrate three benefits offered by the proposed model over traditional topic models, as listed in the introduction. First, we explore how the joint modeling of creative documents and their summaries sheds light on the process by which humans summarize creative documents, and enhances our understanding of the significance of the topics estimated by the model. Second, we show how the model may be used to construct various measures of distinctiveness for creative documents, which can inform the debate on the link between distinctiveness and success in creative industries. Third, we present a proof of concept of an online tool based on the model, which assists humans in writing summaries of creative documents. Throughout the remainder of the paper, we focus on the results based on estimating the model on the calibration sample, in each dataset.

Model Output: Topics and Offset Variables

As mentioned earlier, we set the number of regular topics K to 100, expecting only some of the topics to be "non-flat." Indeed, we find that the number of regular topics that have meaningful variations in their weights $\{\beta_{kv}^{reg}\}_v$ and intensities $\{\theta_{dk}^{reg}\}_d$ is 30 for the marketing academic papers dataset, 24 for the movies dataset, and 19 for the TV shows dataset.[6] These regular topics are not defined merely as groups of words that tend to appear together in documents, but rather as groups of words that tend to appear together in documents *and* that tend to have similar weights in summaries relative to documents. In addition, each topic comes with an offset variable, which quantifies the extent to which the topic was deemed "summary worthy" by the humans who wrote the summaries of the documents in the corpus. Figure 2 plots, for the marketing academic papers dataset, the distribution of the offset variables across the non-flat regular topics, ϵ_k^{reg}, together with the distribution of the offset variables for the residual topics, ϵ_d^{res}, across documents in the calibration sample. We see some variation in offset variables across regular topics, confirming that there is value in allowing each regular topic to be weighed differently in summaries vs. full documents. In particular, two of the regular topics are outliers with very large offset variables, and there appears to be a mass of regular topics with very low offset variables. The corresponding distributions for the other datasets are reported in Web Appendix A. We find that the standard deviation of ϵ_k^{reg} across regular topics is smallest in the marketing academic papers dataset (std=0.16), followed by the movies dataset (std=1.99) and the TV shows dataset (std=11.41). This may be interpreted as suggesting that the difference in content between

[6] We identify non-flat topics based on the standard deviation of topic weights across words. There is always a mass of topics that have very low standard deviation. Because the exact value of this low standard deviation varies slightly across datasets, we do not apply a fixed cutoff, but rather identify the mass of flat topics on a case by case basis, by inspection.

synopses and dialogues is greater than the difference between synopses and scripts, which itself is greater than the difference between academic abstracts and papers, which has good face validity. We note that the introduction of residual topics reduces the number of non-flat regular topics and changes their content. Indeed, the nested version without residual topics finds 100 non-flat topics in all three datasets. In addition, the regular topics identified by the nested version without residual topics have less variation in ϵ_k^{reg}: the standard deviation of ϵ_k^{reg} across regular topics is decreased respectively to 0.01, 1.07 and 2.51 in the marketing academic papers, movies, and TV shows datasets.

Figure 3 reports the distribution of the proportion of fitted content assigned to the residual topic ("outside the cone") in documents and summaries, for the academic papers dataset. The proportion of fitted "outside the cone" content in document d is measured as: $\frac{\sum_v \beta_{dv}^{res}}{\sum_v [\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} + \beta_{dv}^{res}]}$, and the proportion of fitted "outside the cone" content in the summary of document d is measured as: $\frac{\sum_v \beta_{dv}^{res} \epsilon_d^{res}}{\sum_v [\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} \epsilon_k^{reg} + \beta_{dv}^{res} \epsilon_d^{res}]}$. In this dataset, these two proportions have a correlation of 0.66 across documents ($p < 0.01$).

<INSERT FIGURES 2 AND 3 ABOUT HERE>

We report descriptions of the non-flat regular topics, and illustrate the type of insight offered by estimating offset variables for these topics. Web Appendix A reports the offset variables, the average topic intensities across documents, and the words with the highest topic weights, for all non-flat regular topics in each dataset. We also visualize some of these topics, by creating word clouds based on randomly drawing words according to a discrete probability distribution with weights proportional to the topic weights β. The position of words in these word clouds has no meaning, but the size of each word indicates its frequency in the simulated data, i.e., its weight on that topic. Figure 4 shows word clouds for the two

regular topics with the smallest offset variables ϵ_k^{reg} in the marketing academic papers dataset. These are topics that tend to be under-represented in summaries compared to full documents. We see that one of these topics has large weights on words like "participants," "people," and "manipulation." We may interpret this topic as providing details related to experiments. The other topic has larger weights on words like "model," "table," "parameters," "estimates," and may be interpreted as providing details related to data analysis. Figure 5 shows word clouds corresponding to the two regular topics that have the largest offset variables, i.e., that tend to be over-represented in the abstracts of marketing academic papers compared to the full papers. We see that one of them has a disproportionately large weight on the word "find," and the other has a disproportionately large weight on the word "firm." These topics might be interpreted as describing the findings of a paper, and its implications for firms. In sum, the results suggest that when writing abstracts of marketing academic papers, authors tend to emphasize the paper's findings and its implications for firms, and under-weigh details related to data collection and data analysis. Such findings have good face validity.

<INSERT FIGURES 4 AND 5 ABOUT HERE>

Web Appendix A displays similar information for the movies and TV shows datasets. In movies, the topics with the lowest offset variables appear to relate to the setting of various scenes in the movie. In TV shows, the two topics with the smallest offset variables appear to relate to standard dialogues. The topic with the largest offset variable appears to relate to actions (e.g., "gets," "takes," "finds," "comes"), and relationships (e.g., "friends," "family"). The topic with the second largest offset variable appears to relate to the appearance of guest stars and other special events in the episode.

The figures and tables reported in this section illustrate the additional layer of infor-

mation provided by the joint modeling of creative documents and their summaries. Offset variables provide insight into the process by which humans summarize creative documents in a particular domain, and enhance our understanding of the significance of each topic. As noted above, the introduction of residual topics reduces the number of non-flat regular topics estimated by the model. Rare topics that are shared by only a small number of documents are likely to be reflected in residual topics rather than regular topics. Hence, if the goal of a researcher is to identify such rare topics, the version of the model that does not include residual topics may be preferred. The inclusion of residual topics, on the other hand, greatly improves the model's ability to fit and predict the content of documents and summaries, and allows researchers to develop a rich set of distinctiveness measures which may be linked to success. Indeed, when residual topics are not present, two of the distinctiveness measures defined in the next section, become unavailable.

Measuring the Distinctiveness of Creative Documents

There has been some debate in the literature on the relationship between distinctiveness and success in creative industries. In this section, we review some of the empirical studies that have contributed to this debate, and show that the proposed model may be used to estimate various measures of distinctiveness, which may help researchers paint a more nuanced picture of the relationship between the distinctiveness and success of creative documents.

Distinctiveness Measures Based on Proposed Model

We consider three distinctiveness measures, that each relies on different aspects of the model. The first, directly based on Berger and Packard (2018), measures the distinctiveness

of the *combination* of *regular* topics in a document. Given a reference group g, we measure the distinctiveness of a document d with intensities on non-flat regular topics $\{\theta_{dk}^{reg}\}_k$, as: $1 - \Sigma_k \frac{|\theta_{dk}^{reg} - \theta_{gk}^{reg}|}{\theta_{dk}^{reg} + \theta_{gk}^{reg} + 0.0001}$. (Berger and Packard, 2018, use the same equation, based on the topic intensities provided by LDA). We refer to this measure of distinctiveness as "inside the cone distinctiveness," because it is based on how the intensities of a document on the regular topics differ from the average in a reference group. We use the journal in which the paper was published, the genre of the movie,[7] and the TV series to which the episode belongs as the reference groups in our respective datasets.

The second measure, which is new to this paper and which is not available from traditional topic models, is based on the "outside the cone" content in the document. We compute, for each creative document, the proportion of fitted content allocated to the residual topic: $\frac{\sum_v \beta_{dv}^{res}}{\sum_v [\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} + \beta_{dv}^{res}]}$. We refer to this measure as "outside the cone distinctiveness."

The third measure, which is also new to this paper and not available from traditional topic models, is based on how "outside the cone" content is weighted in the document's summary, relative to the full document. We simply consider the offset variable on the document's residual topic, ϵ_d^{res}, as a measure of how strongly the document's summary emphasizes the "outside the cone" content from the document. We refer to this measure as "outside the cone emphasis in summary."

In our analyses, we standardize all three measures across documents, for interpretability. The correlations between the three distinctiveness measures in each dataset are reported in Table 3. The lack of consistently high correlation between any pair of measures suggests that these three measures indeed capture different aspects of creative documents.

[7] If a movie belongs to multiple genres, we average the distinctiveness measure across these genres.

<INSERT TABLE 3 ABOUT HERE>

Distinctiveness vs. Success in Academic Papers

The extant literature has found a general positive relationship between distinctiveness and number of citations in academic papers (Uzzi et al., 2013).[8] We extract the number of citations of each paper in our dataset using the API offered by Crossref (www.crossref.org).[9] For the papers in our calibration dataset, we regress the log of 1 plus the number of citations (we take the log due to the skewness of the number of citations) on the three distinctiveness measures.[10] We control for journal fixed effects, publication year fixed effects, the paper's intensities on (non-flat) regular topics $\{\theta_{dk}^{reg}\}_k$, and the paper's number of pages. Results are provided in Table 4. We see that all three measures of distinctiveness are significantly and positively related to a paper's number of citations. That is, in our data the number of citations received by a marketing academic paper tends to be higher when the paper uses an unusual combination of intensities on regular topics, when the paper features more "outside the cone" content, and when the abstract weighs this content disproportionately. The magnitudes of the regression coefficients suggest that the strongest relation is with "outside the cone distinctiveness" (recall all distinctiveness measures are standardized).[11]

[8] Uzzi et al. (2013) measure distinctiveness based on the combinations of journals cited by the paper.

[9] Based on a random subsample of 100 papers, we find that the number of citations provided by Crossref correlates very highly with Google Scholar ($\rho = 0.964$) and ISI ($\rho = 0.979$), while offering the benefit of being publicly available via an API.

[10] We also tested specifications that include a square term for each measure, to allow for diminishing returns to distinctiveness. However, we find no evidence of significant diminishing returns to distinctiveness in any of our dataset.

[11] We repeated the analysis on a sample of 632 papers published between 2010 and 2015 in top Sociology journals (*European Sociological Review*, *American Sociological Review* and *American Journal of Sociology*; the data were graciously made available to us by Boghrati et al. (2020)). On this smaller dataset, we also find that the strongest relation is with "outside the cone distinctiveness," although the coefficient is only marginally significant. The coefficients for the other two measures are not significant. We hope future research will explore potential commonalities and differences across academic fields.

These results are purely correlational. Moreover, we were not able to include all the variables from all previous analyses of the factors of citations of marketing academic articles (e.g., Stremersch et al., 2007, 2015, who do not focus on distinctiveness). Our goal is not to make definitive claims on the causal relation between distinctiveness and number of citations of marketing academic papers. Rather, our goal is to illustrate how the distinctiveness measures derived from the proposed model may be used by researchers interested in contributing to that literature. Interestingly, we find that at least in this dataset, the content of summaries appears to be related to the success of creative documents. This echoes recent findings by Pryzant et al. (2017), who study the link between the presence of certain phrases in the description of products in e-commerce platforms (e.g., including references to authority or seasonality), and product sales. Given the ubiquity of summaries across creative industries, further research may be conducted that links the success of creative products to variations in the content of their summaries.

<INSERT TABLE 4 ABOUT HERE>

Distinctiveness vs. Success in Entertainment Products

While the extant literature makes a clear prediction on the link between distinctiveness and citations in academic papers, the literature is not as clear on the link between distinctiveness and success in the context of entertainment products. On the one hand, Berger and Packard (2018) show that songs whose lyrics are more different from their genres, are ranked higher in digital downloads. Danescu-Niculescu-Mizil et al. (2012) and Askin and Mauskapf (2017) also find that distinctiveness is an attractive feature of entertainment products. On the other hand, according to Salganik et al. (2006), the content of entertainment products has

little impact on the success of these products. Such claim echoes previous research by Bielby and Bielby (1994), who also report a quote from a past president of CBS entertainment that "all hits are flukes," and Hahn and Bentley (2003).

Ratings. We first analyze the link between the three measures of distinctiveness and the ratings of movies and TV shows. For each movie in our calibration dataset, we collect the average rating from IMDB (based on the ratings of IMDB users), which we standardize across movies for interpretability. We include fixed effects for the movie's MPAA rating, fixed effects for the movie's genre(s), the movie's intensities on the (non-flat) regular topics, the movie's duration (in min), and the log of the movie's production budget (in $, adjusted for inflation, using the tool available at https://data.bls.gov/cgi-bin/cpicalc.pl). All these control variables (with the exception of the intensities on regular topics) are obtained from IMDB. Results are provided in the first column of Table 5. We find that "outside the cone distinctiveness" is positively related to the movie's rating. Interestingly, "inside the cone distinctiveness" is actually negatively related to the movie's rating in our dataset, i.e., movies whose regular topic intensities deviate more from the mean of their genre tend to receive lower ratings. We also find that "outside the cone emphasis in summary" is not significantly related to ratings. This is not surprising, given that the role played by synopses in the movie industry is more restricted than the role played by abstracts in academia.

For TV shows, we were able to obtain IMDB ratings for 9,358 of the episodes in our calibration dataset (some episodes were not found on IMDB, and IMDB reports ratings only for episodes that were rated by at least five users). In our analysis, we only kept episodes from TV series for which we had ratings on at least two episodes, in order to include fixed effects for each TV series. This resulted in 9,285 observations and 318 fixed effects. In addition,

we control again for the episode's intensities on the (non-flat) regular topics. Results are provided in Table 6. In this dataset, consistent with the analysis of movie ratings, we find that "outside the cone distinctiveness" is positively related to the TV show's rating. The coefficients for the other two measures of distinctiveness are not statistically significant.

Return on investment. Finally, for movies, we analyze the link between distinctiveness and financial success, measured as the log of the movie's return on investment, defined as in Eliashberg et al. (2014) as the ratio of the movie's domestic box office performance (also obtained from IMDB) to its production budget. In addition to the controls included in the first regression reported in Table 5, we control for the movie's rating. Results, based again on our calibration dataset, are provided in the second column of Table 5. In this dataset, we find that none of the distinctiveness measure is significantly related to financial success.

<INSERT TABLES 5 AND 6 ABOUT HERE>

Discussion

Our analysis suggests that "inside the cone distinctiveness," "outside the cone distinctiveness," and "outside the cone emphasis in summary" provide meaningful and useful measures of distinctiveness, which may have different relations to success, depending on the context and on how "success" is defined and measured. Across our three datasets, "outside the cone distinctiveness" (which is new to this paper) is robustly and positively associated with success. In contrast, "inside the cone distinctiveness" (which is directly based on extant research) is positively related to the number of citations of marketing academic papers, but negatively related to movie ratings. This is not inconsistent with the literature, which suggests that distinctiveness should be positively related to success for academic papers, but

which is more ambivalent on the link between distinctiveness and success in entertainment industries. Finally, in the context of marketing academic papers, we find that putting more emphasis in the abstract on the "outside the cone" content from the paper, is associated with a larger number of citations.

We note that our measures of distinctiveness are based on the entire set of training documents, and hence do not capture novelty with respect to *contemporaneous* documents. In particular, some documents may have been novel at the time at which they were released/published and may have become influential, leading to similar future documents. Such novel documents may not score high on our distinctiveness measures despite being novel, due to the presence of similar documents in the corpus. The dynamic version of the model described in Web Appendix G addresses this issue by allowing topics to evolve over time, hence measuring the distinctiveness of a document with respect to the topics defined at the time this document was published. We apply this dynamic extension of the model to our marketing academic papers datasets, which contains all papers published in a set of journals over 6 years. We find that "inside the cone distinctiveness," "outside the cone distinctiveness" and "outside the cone emphasis in summary" are still all positively and significantly related to the number of citations, when measured based on the dynamic version of the model.

In Web Appendix H, we also test various alternative measures of distinctiveness, and alternative ways to explore the link between distinctiveness and success. We find that as the vocabulary size changes, the significance of some of the coefficients associated with distinctiveness measures may change, although we see no reversal (i.e., a coefficient that is significant in one direction under one vocabulary size is never significant in the other direction

under a different vocabulary size). We conduct a simulation study, which illustrates how measures of distinctiveness are affected as relevant words are omitted from the vocabulary or as irrelevant words are included in the vocabulary. This simulation study confirms that the selection of the vocabulary size is bound to have some impact on the output of the model. While this is not an attractive feature, unfortunately this is a characteristic of any topic model, not just the one presented here. We test alternative specifications that link distinctiveness to financial success in the movies dataset, yielding similar results to those reported in Table 5. We measure "inside the cone distinctiveness" using the entire set of training documents as the reference group, rather than documents in the same journal / genre / TV series, and find similar results as those reported in Tables 4 to 6. We perform an analysis that reflects the fact that measures of distinctiveness are constructed from model parameters that are estimated with uncertainty rather than measured precisely. We run each regression 1,000 times using different draws from the posterior distribution of the model parameters, and report the average coefficients as well as whether the 90% and 95% credible intervals include 0. Results are consistent with those reported in Tables 4 to 6. Finally, we measure distinctiveness using standard topic models (LDA and Poisson Factorization), rather than the proposed model. "Inside the cone distinctiveness" is the only distinctiveness measure available from these models, and we find that it is never statistically significantly related to success in any of the regression, with the exception of "inside the cone distinctiveness" estimated based on the standard Poisson Factorization which is marginally related to return of investment of movies.

Computer-Assisted Summary Writing

As mentioned in our literature review, the traditional approach in the computer science literature would be to attempt to completely automate the summarization of documents, typically via sentence extraction. We argue that this approach is less relevant in the context of *creative* documents. In particular, the nature of creative documents is such that the stakes are usually high enough for humans to be motivated and available to write summaries. For example, the author or publisher of a new book typically has enough motivation to write a synopsis for this document, and may not find as much value in a tool that would automatically generate a summary. Similar comments may be made about the publisher of a new movie or play, the author of an academic paper, the developer of an innovative product, the author of a business plan, etc. This is in contrast to the traditional text summarization literature that typically deals with the summarization of large volumes of documents such as news articles, where automation has significant cost saving implications. Moreover, sentence extraction is likely to be an inappropriate text summarization approach in many creative contexts. For example, an abstract of a scientific paper made exclusively of sentences from the paper, or a TV show synopsis made exclusively of sentences from the show's dialogues, may be unacceptable to the relevant audience. Accordingly, we argue that in our situation, it is more useful to develop decision support tools that assist humans in writing summaries of creative documents, rather than developing automatic text summarization tools featuring sentence extraction.

We have built a proof of concept for such a decision support tool, using php and a mysql database. The tool allows a user to upload a creative document that was not necessarily part of the corpus on which the model was estimated. When the user submits a new document

d^{out}, the text of this document is tokenized on the fly (using custom-built php code developed by the authors), and the number of occurrences of each word in the vocabulary is computed for that document. Intensities on the regular topics $\{\theta^{reg}_{d^{out}k}\}_k$ and the residual topic $\beta^{res}_{d^{out}}$ for the new document d^{out} are estimated in real time using variational inference, given the other model parameters.[12]

As output, the tool reports representative words for the five regular topics with the largest intensities ($\theta^{reg}_{d^{out}k}$) and representative words for the residual topic. In addition, the tool reports representative words for the five regular topics that the model predicts should have the largest intensities in the summary of the new document (i.e., the regular topics with the largest values of $\theta^{reg}_{d^{out}k} \times \epsilon^{reg}_k$). In our current implementation, for each topic we report the 10 words with the highest weights on the topic (β_{kv}) as representative words.[13]

Because our model should be run separately in each domain, we customize the tool for each domain of application. We have created one version of our online tool corresponding to each corpus studied in this paper (marketing academic papers, movies, and TV shows). This proof of concept is publicly available at http://creativesummary.org.[14]

Importantly, such decision support tool may also leverage analysis such as the one re-

[12]Variational inference is performed on the fly within php, using code developed by the authors. In order to speed up computations, the di-gamma function $\Psi(x)$ is approximated as follows. If $x < 2$, x is rounded to the nearest thousandth, and $\Psi(x)$ is obtained from a look-up table. If $x \geq 2$, $\Psi(x)$ is approximated by its asymptotical expansion, with precision $O(\frac{1}{x^{16}})$. Also in order to speed up computations, the ascent mean-field variational inference algorithm is run for 100 iterations systematically, rather than checking convergence at each iteration. These approximations are made only in the online tool. With the current implementation, all computations for a new document are typically performed within 5 seconds.

[13]We only show words that have sufficient weights on the topic: $\frac{\beta_{kv}}{\sum_{v'} \beta_{kv'}} > 0.01$. If the topic is "flat" and no word satisfies this criterion, we do not report the topic.

[14]The php code is common across domains. The vocabulary for each domain, the weights of the regular topics $\{\beta^{reg}_k\}_k$, and the offset variables on regular topics $\{\epsilon^{reg}_k\}_k$, which are all obtained from estimating the model on the corresponding corpus, are stored in the database that supports the tool. Creating a version of the tool for a new domain (e.g., business plans) only requires running the model on a corpus from this domain, and uploading the results onto the database.

ported in the previous section , in order to help the user improve the effectiveness of their summary. For example, we found that marketing academic papers in which the abstract puts more emphasis on the "outside the cone" content in the paper (i.e., higher ϵ_d^{res} or "outside the cone emphasis in summary"), tend to have more citations. Without making unfounded causal claims, we can report this correlational finding to the user of the online tool. Accordingly, in the proof of concept of the tool tailored for marketing academic papers, we include the following statement next to the representative words from the "outside the cone" topic: "Our research suggests that a paper whose abstract puts more emphasis on the paper's 'outside the cone' topic, tends to receive more citations."

Conclusions

Our contribution in this paper is primarily methodological. We develop and apply a new topic model designed specifically for the study of creative documents. Guided by the creativity literature, this model nests and extends Poisson Factorization in two ways. First, we explicitly model residual, "outside the cone" content and how it is represented in summaries vs. documents. Second, we jointly model the content of documents and their summaries, and quantify (using offset variables) how the intensity of each topic differs systematically in summaries compared to full documents. We validate the model using three different datasets containing marketing academic papers (summarized by abstracts), movie scripts (summarized by synopses), and TV show closed captions (summarized by synopses). The proposed model offers the standard benefits of topic models, i.e., it extracts topics from a corpus of documents, and assigns intensities on each topic for each document (although the introduction of residual topics changes the number and content of the non-flat regular topics).

We illustrate three additional benefits provided by our model for academics and practitioners. First, the offset variables estimated by our model, which quantify the extent to which each topic was deemed "summary worthy" by the humans who wrote the summaries of the documents in the corpus, shed light into the process by which humans summarize creative documents and help understand the significance of each topic. Second, we illustrate how the model may be used to construct new measures of distinctiveness for creative documents, which have the potential to shed new light on the relation between distinctiveness and success in creative industries. Third, we develop an online, interactive, freely accessible tool based on our model, which provides a proof of concept for using the model's output to assist humans in writing summaries of creative documents.

We close by highlighting additional areas for future research. First, it would be interesting to introduce covariates into the model, that influence the topic intensities and/or the offset variables. In the context of entertainment products, such covariates might include genres, country of origin, etc. In the context of academic papers, these may include subfields, whether the paper is based on a dissertation, etc. Second, alternative topic models may capture the *structure* of creative documents (e.g., different sections, scenes, acts). Third, it would be interesting to study how the content of summaries varies systematically based on the objectives of the summary. For example, in some cases summaries serve primarily as "teasers" for creative products, while in others they serve more as "substitutes" for the products. For example, the offset variables might differ systematically between spoilers and synopses, or between abstracts written for conferences vs. journal articles.

Tables

Table 1: Descriptive Statistics.

	Metric	Unit of analysis	Mean	St. dev.	Range
Marketing academic papers	Number of word occurrences	Paper	2110.26	647.31	[12;5016]
	Number of word occurrences	Abstract	41.39	15.15	[4;125]
	Number of words with at least one occurrence	Paper	269.74	56.54	[7;409]
	Number of words with at least one occurrence	Abstract	23.44	7.56	[3;61]
	Number of occurrences across full texts	Word	2812.97	4020.34	[188;44091]
	Number of occurrences across abstracts	Word	55.18	98.96	[0;1420]
	Number of full texts with at least one occurrence	Word	359.57	268.50	[1;1216]
	Number of abstracts with at least one occurrence	Word	31.25	48.00	[0;624]
Movies	Number of word occurrences	Script	1490.86	580.36	[0;7489]
	Number of word occurrences	Synopsis	91.26	80.93	[1;748]
	Number of words with at least one occurrence	Script	310.17	69.05	[0;533]
	Number of words with at least one occurrence	Synopsis	46.78	31.18	[1;219]
	Number of occurrences across scripts	Word	1279.16	2016.88	[89;33633]
	Number of occurrences across synopses	Word	78.30	122.36	[0;1322]
	Number of scripts with at least one occurrence	Word	266.12	196.20	[1;834]
	Number of synopses with at least one occurrence	Word	40.14	52.26	[0;426]
TV show episodes	Number of word occurrences	Closed caption	797.77	405.76	[0;3819]
	Number of word occurrences	Synopsis	8.20	10.76	[0;261]
	Number of words with at least one occurrence	Closed caption	289.70	80.95	[0;718]
	Number of words with at least one occurrence	Synopsis	7.35	7.39	[0;155]
	Number of occurrences across closed captions	Word	21189.56	41049.97	[1469;693406]
	Number of occurrences across synopses	Word	217.72	283.79	[0;2889]
	Number of closed captions with at least one occurrence	Word	7694.59	5526.92	[24;26148]
	Number of synopses with at least one occurrence	Word	195.22	248.20	[0;2498]

Note: there are 1,333 papers in the academic papers dataset, 858 movies in the movies dataset, and 26,561 TV show episodes in the TV shows dataset. There are 1,000 words in the vocabulary for each dataset. The first column (vertically aligned) contains the dataset, the second the metric of interest, the third the unit of analysis, and the remaining columns report the mean, standard deviation, min and max of the correspond metric across the units of analysis. For example, the first row indicates that in the marketing academic papers dataset, papers have on average 2,110.26 word occurrences.

Table 2: Fit and Predictive Performance.

		Fit			Predictive perf.
		Calibration documents	**Calibration summaries**	**Validation documents**	**Validation summaries**
Marketing academic papers	Full Model	104.54	67.60	109.89	82.50
	No residual topic	197.04	141.39	227.73	169.90
	ϵ constant	104.46	73.13	109.92	85.74
	No residual topic and ϵ constant	196.91	146.39	227.54	176.83
	Residual topics only	101.83	70.84	106.68	84.30
	LDA	197.13	145.73	227.12	176.34
Movies	Full Model	168.72	177.28	179.11	290.04
	No residual topic	265.80	279.76	324.51	358.17
	ϵ constant	169.07	213.13	179.25	348.25
	No residual topic and ϵ constant	265.50	346.99	324.54	428.55
	Residual topics only	163.82	204.87	172.39	355.07
	LDA	267.29	343.28	328.05	424.35
TV show episodes	Full Model	241.61	246.36	241.08	410.90
	No residual topic	361.10	416.28	358.90	439.27
	ϵ constant	241.58	311.23	240.85	577.77
	No residual topic and ϵ constant	360.29	633.89	358.39	686.96
	Residual topics only	234.86	294.33	233.68	574.21
	LDA	360.56	643.63	358.33	696.76

Note: fit and predictive performance are measured using perplexity (lower values indicate better fit).

Table 3: Correlation Between Distinctiveness Measures.

		"Outside the cone distinctiveness"	"Outside the cone emphasis in summary"
Marketing academic papers	"Inside the cone distinctiveness"	-0.03	-0.16**
	"Outside the cone distinctiveness"		0.15**
Movies	"Inside the cone distinctiveness"	-0.48**	-0.14**
	"Outside the cone distinctiveness"		0.05
TV shows	"Inside the cone distinctiveness"	0.25**	0.03**
	"Outside the cone distinctiveness"		0.09**

Note: *: significant at $p < 0.10$. **: significant at $p < 0.05$.

Table 4: Link Between Distinctiveness Measures and Citations - Marketing Academic Papers.

Covariates	**DV=log(1+#citations)**
Journal fixed effects	✓
Publication year fixed effects	✓
Intensities of paper on non-flat regular topics	✓
Number of pages in paper	0.040**
"Inside the cone distinctiveness" (from journal)	0.113**
"Outside the cone distinctiveness"	0.140**
"Outside the cone emphasis in summary"	0.059**
Number of parameters	43
Number of observations	1,000
R^2	0.353

Note: OLS regression. *: significant at $p < 0.10$. **: significant at $p < 0.05$. All three distinctiveness measures are standardized across papers for interpretability.

Table 5: Link Between Distinctiveness Measures and Performance - Movies.

Covariates	**DV=movie rating**	**DV=log(return on investment)**
MPAA rating fixed effects	✓	✓
Genre fixed effects	✓	✓
Intensities of script on non-flat regular topics	✓	✓
Movie duration (in min)	0.003*	-0.003
Log(inflation-adjusted production budget)	-0.093**	-0.329**
Movie rating	–	0.451**
"Inside the cone distinctiveness" (from genre)	-0.090**	-0.027
"Outside the cone distinctiveness"	0.253**	-0.109
"Outside the cone emphasis in summary"	0.050	0.069
Number of parameters	54	55
Number of observations	596	581
R^2	0.357	0.262

Note: each column corresponds to one regression estimated separately using OLS. *: significant at $p < 0.10$. **: significant at $p < 0.05$. All three distinctiveness measures and movie ratings are standardized across movies for interpretability. Observations in the first (resp., second) regression are limited to movies for which production budget was available (resp., production budget and box office performance were available).

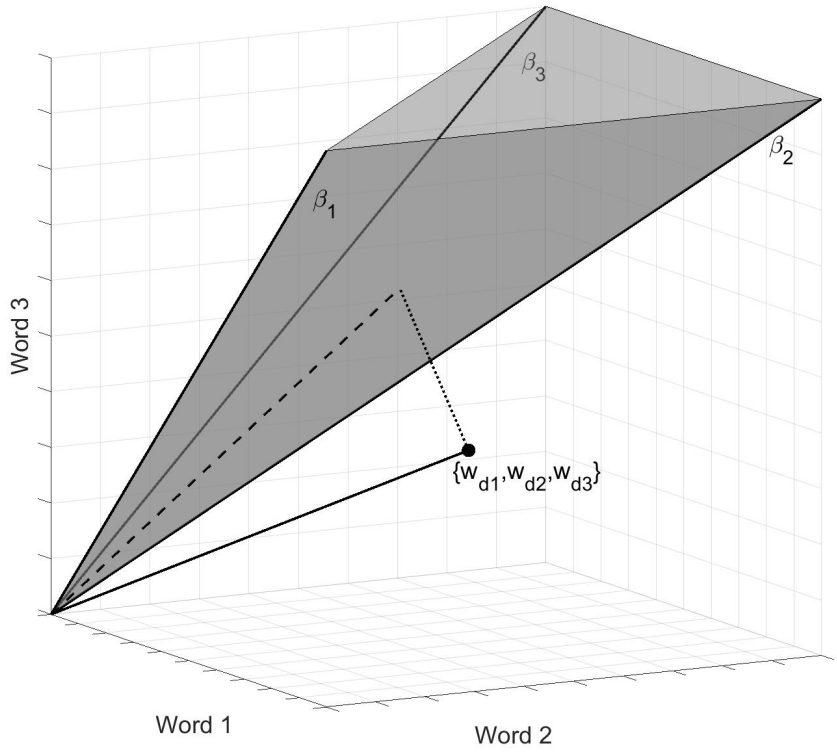
Table 6: Link Between Distinctiveness Measures and Performance - TV Episodes.

Covariates	**DV=episode rating**
TV series fixed effects	✓
Intensities of script on non-flat regular topics	✓
"Inside the cone distinctiveness" (from TV series)	-0.005
"Outside the cone distinctiveness"	0.074**
"Outside the cone emphasis in summary"	-0.008
Number of parameters	340
Number of observations	9,285
R^2	0.687

Note: OLS regression. *: significant at $p < 0.10$. **: significant at $p < 0.05$. All three distinctiveness measures and episode ratings are standardized across episodes for interpretability.

Figures

Figure 1: Geometric Interpretation of "Inside the Cone" vs. "Outside the Cone" Content.



Note: in this example with three words in the vocabulary and three topics, each vector β_k represents the weights of each word on topic k. The grey cone contains all positive combinations of the three topics. The black dot represents a vector that contains the number of occurrences of each word in a document d: $\{w_{d1}, w_{d2}, w_{d3}\}$. The dashed line represents the projection of this vector on the cone defined by the three topics ("inside the cone" content, captured by standard Poisson Factorization). The dotted line represents the residual ("outside the cone" content, captured by the proposed model but not by standard Poisson Factorization).

Figure 2: Distribution of Offset Variables - Marketing Academic Papers.

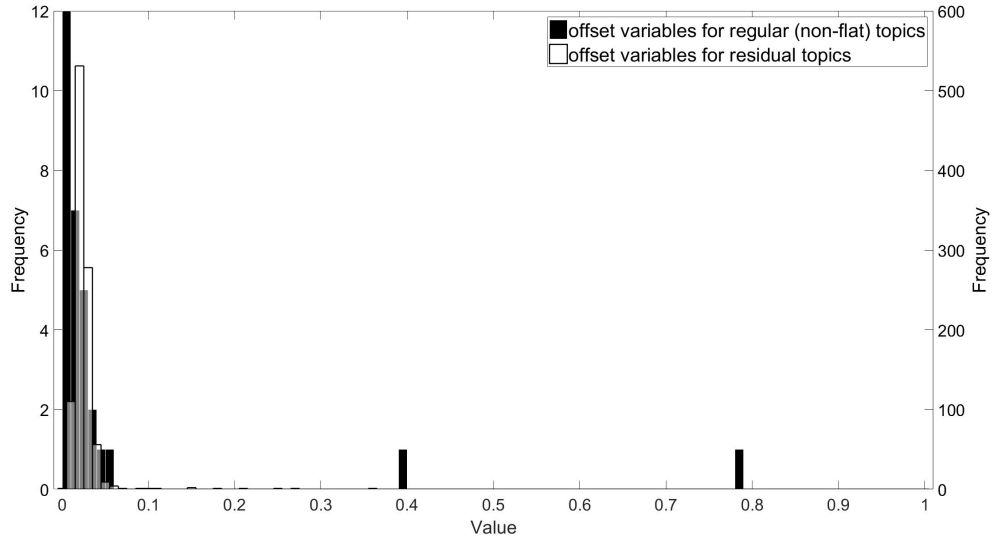
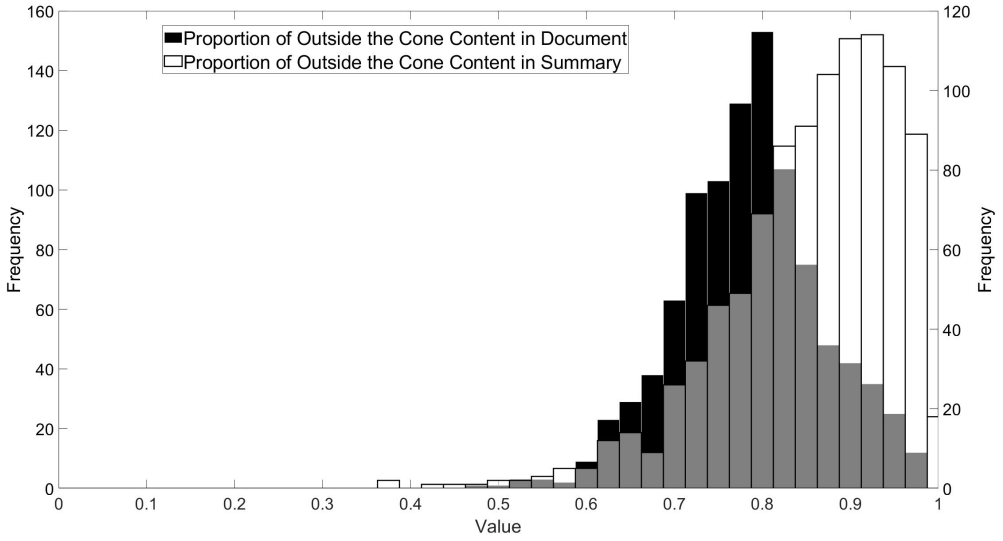
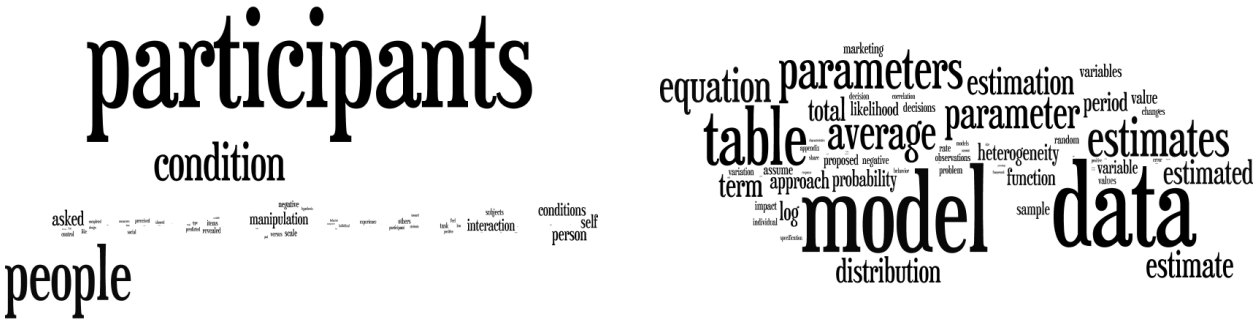


Figure 3: Distribution of the Proportion of fitted "Outside the Cone" Content in Documents and Summaries - Marketing Academic Papers.



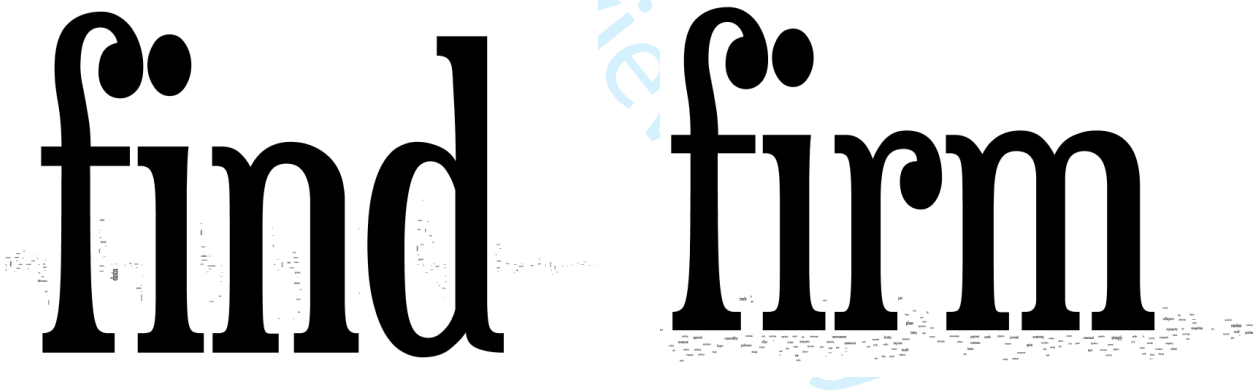
Note: the proportion of fitted "outside the cone" content in document d is measured as: $\frac{\sum_v \beta_{dv}^{res}}{\sum_v [\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} + \beta_{dv}^{res}]}$. The proportion of fitted "outside the cone" content in the summary of document d is measured as: $\frac{\sum_v \beta_{dv}^{res} \epsilon_d^{res}}{\sum_v [\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} \epsilon_k^{reg} + \beta_{dv}^{res} \epsilon_d^{res}]}$. The correlation between the two proportions is 0.66 across documents in the calibration sample ($p < 0.01$).

Figure 4: Word Clouds for Regular Topics with Smallest Offset Variables - Marketing Academic Papers.



Note: these topics tend to be under-represented in summaries compared to full texts.

Figure 5: Word Clouds for Regular Topics with Largest Offset Variables - Marketing Academic Papers.



Note: these topics tend to be over-represented in summaries compared to full texts.

References

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*, 2017.

Noah Askin and Michael Mauskapf. What makes popular culture popular? product features and optimal differentiation in music. *American Sociological Review*, 82(5):910–944, 2017.

Jonah Berger and Grant Packard. Are atypical things more popular? *Psychological science*, 29(7):1178–1184, 2018.

William T Bielby and Denise D Bielby. " all hits are flukes": Institutionalized decision making and the rhetoric of network prime-time program development. *American Journal of Sociology*, 99(5):1287–1313, 1994.

David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

David M Blei and John D Lafferty. Topic models. *Text mining: classification, clustering, and applications*, 10(71):34, 2009.

David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent dirichlet allocation. *Machine Learning*, 3(4/5):993–1022, 2003.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *arXiv preprint arXiv:1601.00670*, 2016.

Reihane Boghrati, Jonah Berger, and Grant Packard. How writing style shapes the impact of scientific findings. *working paper*, 2020.

Joachim Büschken and Greg M Allenby. Sentence-based text analysis for customer reviews. *Marketing Science*, 35(6):953–975, 2016.

John Canny. Gap: a factor model for discrete data. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 122–129. ACM, 2004.

Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. You had me at hello: How phrasing affects memorability. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 892–901. Association for Computational Linguistics, 2012.

Hal Daumé III and Daniel Marcu. Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 305–312. Association for Computational Linguistics, 2006.

Jean-Yves Delort and Enrique Alfonseca. Dualsum: a topic-model based approach for update summarization. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 214–223. Association for Computational Linguistics, 2012.

Joshua Eliashberg, Sam K. Hui, and John Z. Zhang. From story line to box office: A new approach for green-lighting movie scripts. *Management Science*, 53(6):881–893, 2007.

Joshua Eliashberg, Sam K. Hui, and Z. John Zhang. Assessing box office performance using

movie scripts: A kernel-based approach. *IEEE Transactions on Knowledge and Data Engineering*, 26(11):2639–2648, 2014.

Ronald A Finke, Thomas B Ward, and Steven M Smith. Creative cognition: Theory, research, and applications. 1992.

Richard Florida. *The Rise of the Creative Class–Revisited: Revised and Expanded*. Basic books, 2014.

Rachel Giora. *On our mind: Salience, context, and figurative language*. Oxford University Press, 2003.

Prem Gopalan, Jake M Hofman, and David M Blei. Scalable recommendation with poisson factorization. *arXiv preprint arXiv:1311.1704*, 2013.

Prem K Gopalan, Laurent Charlin, and David Blei. Content-based recommendations with poisson factorization. In *Advances in Neural Information Processing Systems*, pages 3176–3184, 2014.

Aria Haghighi and Lucy Vanderwende. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370. Association for Computational Linguistics, 2009.

Matthew W Hahn and R Alexander Bentley. Drift as a mechanism for cultural change: an example from baby names. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(Suppl 1):S120–S123, 2003.

Jia Liu and Olivier Toubia. A semantic approach for estimating consumer content preferences from online search queries. *Marketing Science (forthcoming)*, 2018.

Kathleen McKeown and Dragomir R Radev. Generating summaries of multiple news articles. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82. ACM, 1995.

Sarnoff Mednick. The associative basis of the creative process. *Psychological review*, 69(3): 220, 1962.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*, 2017.

Ani Nenkova and Kathleen McKeown. A survey of text summarization techniques. In *Mining text data*, pages 43–76. Springer, 2012.

Reid Pryzant, Youngjoo Chung, and Dan Jurafsky. Predicting sales from the language of product descriptions. In *eCOM@ SIGIR*, 2017.

Dinesh Puranam, Vishal Narayan, and Vrinda Kadiyali. The effect of calorie posting regulation on consumer opinion: A flexible latent dirichlet allocation model with informative priors. *Marketing Science*, 36(5):726–746, 2017.

Dragomir R Radev and Kathleen R McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):470–500, 1998.

Dragomir R Radev, Eduard Hovy, and Kathleen McKeown. Introduction to the special issue

on summarization. *Computational linguistics*, 28(4):399–408, 2002.

Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856, 2006.

Statista. Average time spent with major media per day in the united states as of april 2016 (in minutes). *https://www.statista.com/statistics/276683/media-use-in-the-us/*, 2017.

Statista. Value of the global entertainment and media market from 2011 to 2021 (in trillion u.s. dollars). *https://www.statista.com/statistics/237749/value-of-the-global-entertainment-and-media-market/*, 2018.

Stefan Stremersch, Isabel Verniers, and Peter C Verhoef. The quest for citations: Drivers of article impact. *Journal of Marketing*, 71(3):171–193, 2007.

Stefan Stremersch, Nuno Camacho, Sofie Vanneste, and Isabel Verniers. Unraveling scientific impact: Citation types in marketing journals. *International Journal of Research in Marketing*, 32(1):64–77, 2015.

Seshadri Tirunillai and Gerard J. Tellis. Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, 51:463–479, August 2014.

Olivier Toubia and Oded Netzer. Idea generation, creativity, and prototypicality. *Marketing Science*, 36(1):1–20, 2017.

Olivier Toubia, Garud Iyengar, Renée Bunnell, and Alain Lemaire. Extracting features of entertainment products: A guided lda approach informed by the psychology of media consumption. *Journal of Marketing Research (forthcoming)*, 2018.

Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

Thomas B Ward. Whats old about new ideas. *The creative cognition approach*, pages 157–178, 1995.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. Recent advances in document summarization. *Knowledge and Information Systems*, 53(2):297–336, 2017.

Ning Zhong and David A. Schweidel. Capturing changes in social media content: A multiple latent changepoint topic model. *working paper*, 2018.

A Poisson Factorization Topic Model for the Study of Creative Documents (and their Summaries): Web Appendix

A Additional Tables and Figures

Table WA1: Non-Flat Regular Topics - Marketing Academic Papers.

Offset variable ϵ_k^{reg}	**Av. topic intensity** θ_{dk}^{reg}	**Words with largest weights**
0.786	0.027	find, data
0.398	0.022	firm
0.054	0.195	consumers, consumer, data, behavior, individual
0.041	0.215	consumers, consumer, conditions, marketing, participants
0.038	0.179	model, data, models, marketing, approach
0.034	0.099	consumer, purchase, decision
0.027	0.257	product, products, positive, impact, consumer
0.024	0.236	marketing, firms, firm, variables, value
0.023	0.268	customer, customers, firm, relationship, firms
0.022	0.274	products, product, category, brand, categories
0.020	0.173	product, participants, products, low, design
0.019	0.157	data, table, models, individual, choice
0.018	0.259	model, models, function, value, product
0.018	0.283	market, firms, data, value, sales
0.016	0.195	consumers, consumer, value, purchase, low
0.015	0.202	price, prices, value, impact, purchase
0.013	0.264	price, prices, consumers, market, model
0.011	0.084	consumers, value, positive
0.008	0.147	low, condition, conditions, variable, average
0.001	0.032	value
<0.001	0.086	data, variable, control, average, regression
<0.001	0.117	interaction, people, control, participants, relationship
<0.001	0.146	positive, behavior, interaction, relationship, low
<0.001	0.148	items, item, scale, measures, marketing
<0.001	0.213	positive, participants, negative, measures, interaction
<0.001	0.309	condition, participants, choice, conditions, asked
<0.001	0.249	participants, consumers, perceived, interaction, manipulation
<0.001	0.285	variables, table, positive, negative, data
<0.001	0.337	participants, people, condition, conditions, asked
<0.001	0.329	model, data, table, parameters, estimates

Note: topics are sorted in decreasing order of the offset variable, reported in the first column. The second column reports the average intensity of that topic across documents. Only words with weights β_{kv}^{reg} larger than 1% of the sum of the weights for that topic ($\sum_v \beta_{kv}^{reg}$), are included in the table.

Table WA2: Non-Flat Regular Topics - Movies.

Offset variable ϵ_k^{reg}	**Av. topic intensity** θ_{dk}^{reg}	**Words with largest weights**
7.783	0.031	orders, killed, attack, several, crew
4.971	0.029	police, killed, building, apartment, several
4.175	0.012	begins, party, film, movie, scene
2.468	0.005	father, son, killed, parents, farm
1.206	0.001	drug, york, dealer, million, lawyer
0.840	0.003	team, major, final, power
0.468	0.021	money, police, gun, killed, murder
0.001	0.003	spins, swings, hey, beat, killed
<0.001	0.097	night, begins, stairs, somebody, hey
<0.001	0.019	house, cut, road, kitchen
<0.001	0.024	beat, book, bathroom, center, game
<0.001	0.028	gonna, em, everybody, son, beat
<0.001	0.026	cont, beat, shit, towards, street
<0.001	0.052	cut, gonna, gun
<0.001	0.042	cut, control, spins, gun, flying
<0.001	0.068	gun, station, blood, police, metal
<0.001	0.090	street, crowd, music, beside, father
<0.001	0.075	cont, hey, begins, gonna, huh
<0.001	0.13188	shit, guy, fucking, cut, fuck
<0.001	0.132	night, bedroom, bathroom, street, kitchen
<0.001	0.173	night, street, hey, cut, guy
<0.001	0.189	gonna, hey, house, shit, night
<0.001	0.179	night, blood, others, begins, cut
<0.001	0.213	street, cars, car, building, police

Note: topics are sorted in decreasing order of the offset variable, reported in the first column. The second column reports the average intensity of that topic across documents. Only words with weights β_{kv}^{reg} larger than 1% of the sum of the weights for that topic ($\sum_v \beta_{kv}^{reg}$), are included in the table.

Table WA3: Non-Flat Regular Topics - TV Shows.

Offset variable ϵ_k^{reg}	**Av. topic intensity** θ_{dk}^{reg}	**Words with largest weights**
49.253	0.950	gets, find, takes, help, wants
10.123	0.950	guest, show, special, join, stars
<0.001	0.950	know, need, really, help, didn
<0.001	0.950	gonna, okay, thank, ready, start
<0.001	0.950	guys, really, thank, great, yeah
<0.001	0.950	yeah, hey, know
<0.001	0.950	people, great, really, much, lot
<0.001	0.950	okay, really, yes, thank, great
<0.001	0.950	like, want, need, okay, put
<0.001	0.950	know, like, want, never, night
<0.001	0.951	gonna, want, cause, talk, life
<0.001	0.951	okay, know, yeah, hey, like
<0.001	0.951	gonna, like, want, really, know
<0.001	0.952	know, like, really, mean, want
<0.001	0.952	hey, yeah, know, gonna, want
<0.001	0.951	know, like, want, something, would
<0.001	0.953	yeah, gonna, know, like, okay
<0.001	0.955	know, like, would, want, yeah
<0.001	0.955	like, know, yeah, gonna, really

Note: topics are sorted in decreasing order of the offset variable, reported in the first column. The second column reports the average intensity of that topic across documents. Only words with weights β_{kv}^{reg} larger than 1% of the sum of the weights for that topic ($\sum_v \beta_{kv}^{reg}$), are included in the table.

Figure WA1: Distribution of Offset Variables - Movies.

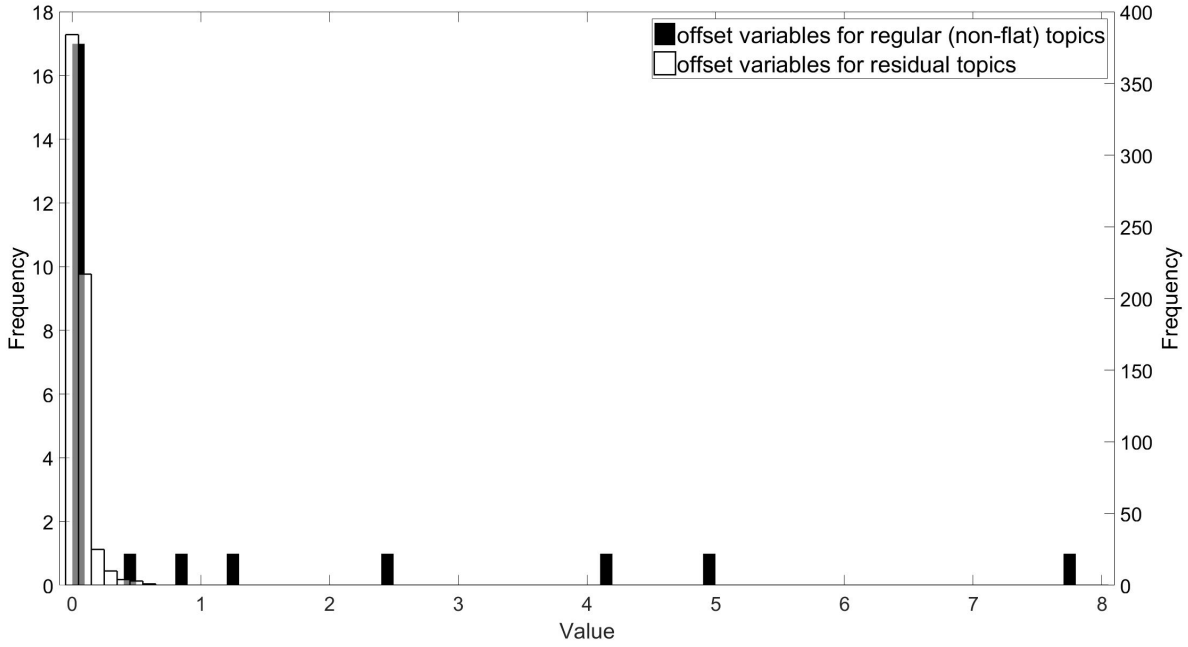
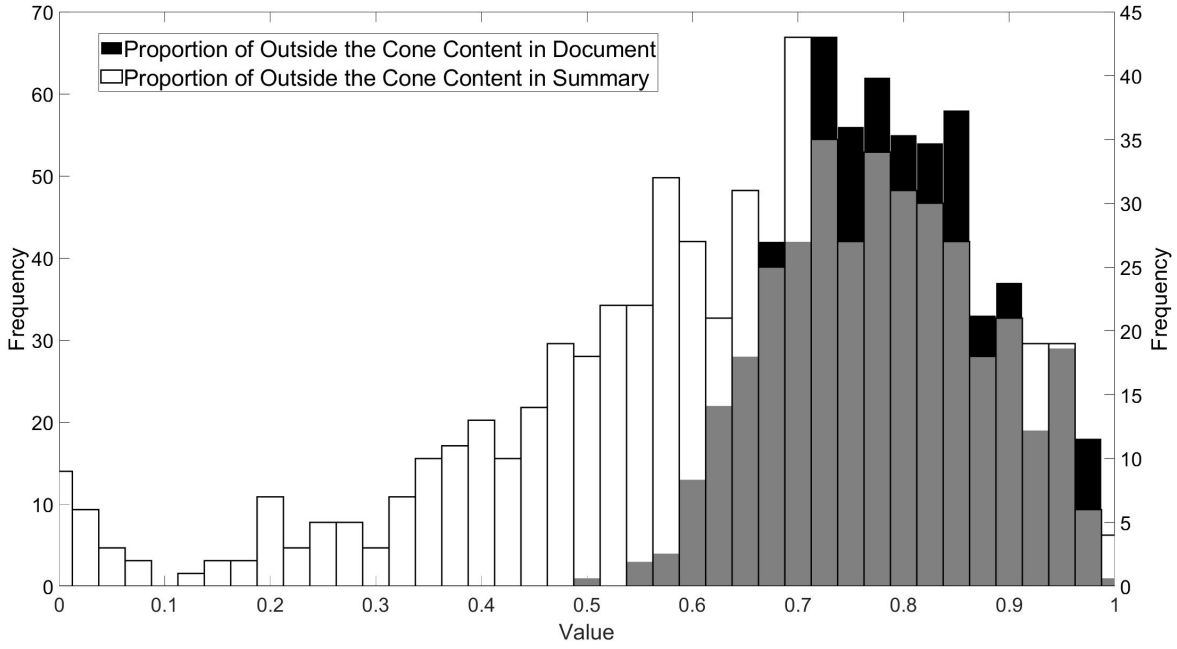


Figure WA2: Distribution of the Proportion of fitted "Outside the Cone" Content in Documents and Summaries - Movies.



Note: the proportion of fitted "outside the cone" content in document d is measured as: $\frac{\sum_v \beta_{dv}^{res}}{\sum_v [\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} + \beta_{dv}^{res}]}$. The proportion of fitted "outside the cone" content in the summary of document d is measured as: $\frac{\sum_v \beta_{dv}^{res} \epsilon_d^{res}}{\sum_v [\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} \epsilon_k^{reg} + \beta_{dv}^{res} \epsilon_d^{res}]}$. The correlation between the two proportions is 0.23 across documents in the calibration sample ($p < 0.01$).

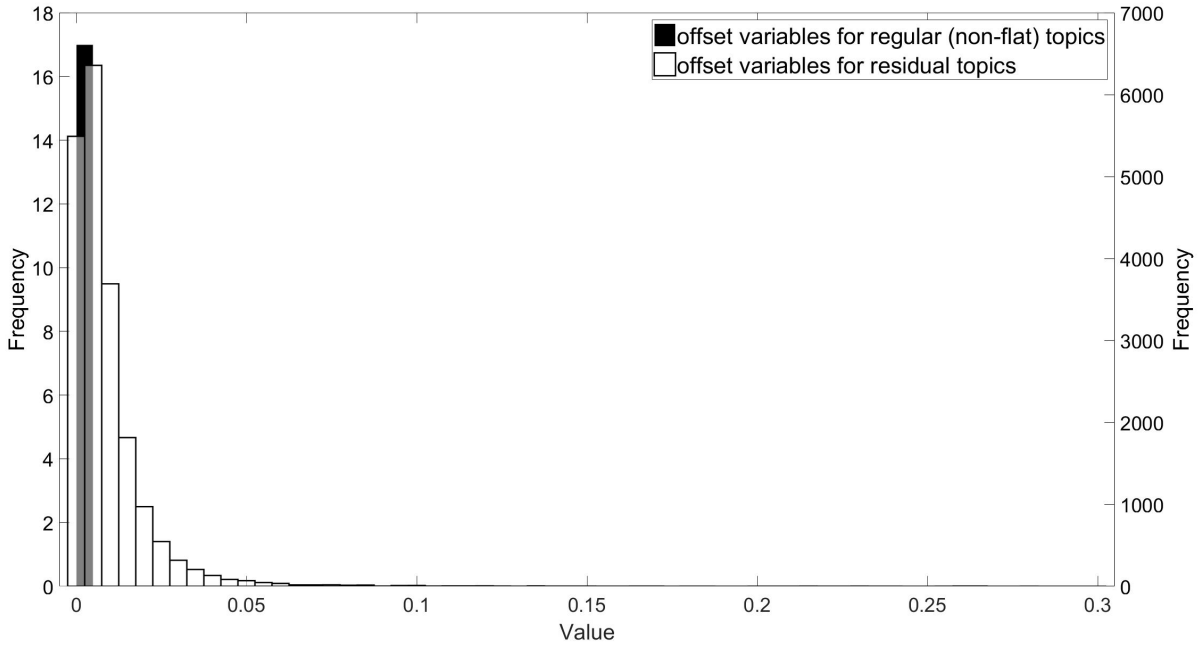
Figure WA3: Word Clouds for Regular Topics with Smallest Offset Variables - Movies.

Note: these topics tend to be under-represented in summaries compared to full texts.

Figure WA4: Word Clouds for Regular Topics with Largest Offset Variables - Movies.

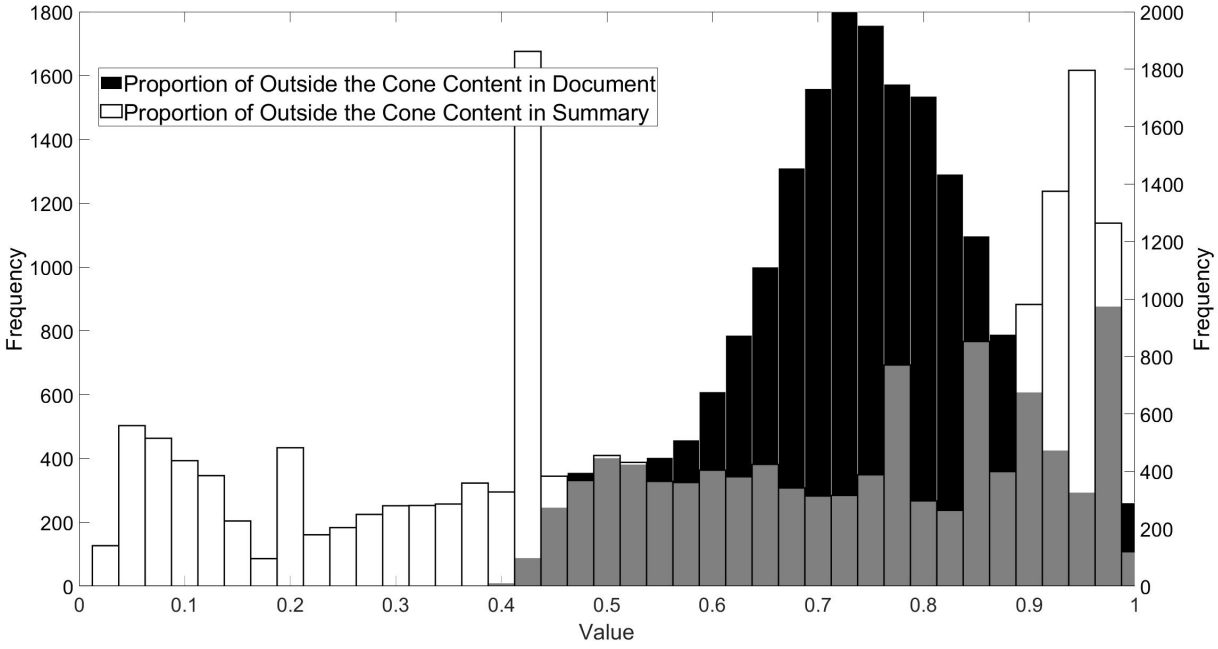
Note: these topics tend to be over-represented in summaries compared to full texts.

Figure WA5: Distribution of Offset Variables - TV Shows.



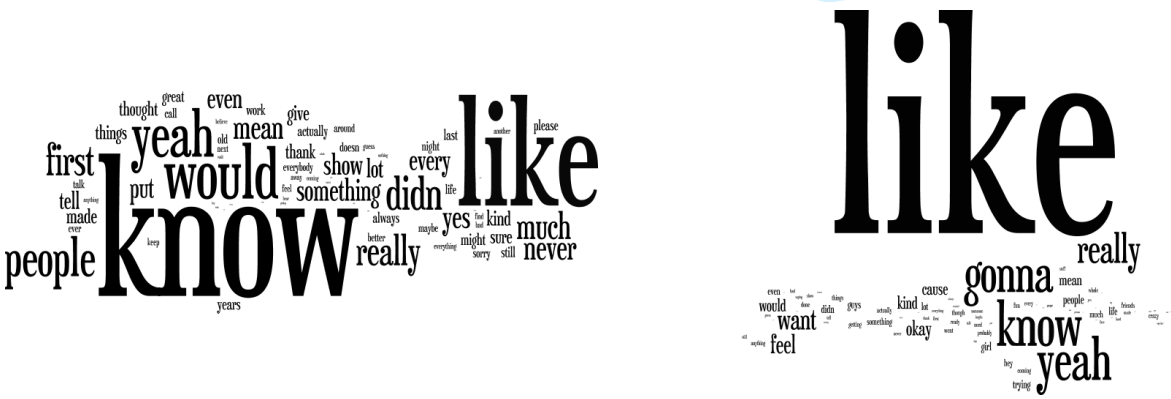
Note: the histogram for the regular topics excludes two outliers with offset variables equal to $\epsilon_k^{reg} = 49.25$ and 10.12.

Figure WA6: Distribution of the Proportion of fitted "Outside the Cone" Content in Documents and Summaries - TV shows.



Note: the proportion of fitted "outside the cone" content in document d is measured as: $\frac{\sum_v \beta_{dv}^{res}}{\sum_v [\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} + \beta_{dv}^{res}]}$. The proportion of fitted "outside the cone" content in the summary of document d is measured as: $\frac{\sum_v \beta_{dv}^{res} \epsilon_d^{res}}{\sum_v [\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} \epsilon_k^{reg} + \beta_{dv}^{res} \epsilon_d^{res}]}$. The correlation between the two proportions is 0.10 across documents in the calibration sample ($p < 0.01$).

Figure WA7: Word Clouds for Regular Topics with Smallest Offset Variables - TV Shows.



Note: these topics tend to be under-represented in summaries compared to full texts.

Figure WA8: Word Clouds for Regular Topics with Largest Offset Variables - TV Shows.

Note: these topics tend to be over-represented in summaries compared to full texts.

B Variational Inference for Model Estimation

The conditional posterior distributions are given as follows (based on Gopalan et al., 2013, 2014):

$$\theta_{dk}^{reg} \sim Gamma(\alpha_3 + \sum_v (z_{dv,k}^{reg} + z_{dv,k}^{sum,reg}), \alpha_4 + \sum_v \beta_{vk}^{reg}(1 + \epsilon_k^{reg}))$$

$$\beta_{kv}^{reg} \sim Gamma(\alpha_1 + \sum_d (z_{dv,k}^{reg} + z_{dv,k}^{sum,reg}), \alpha_2 + \sum_d \theta_{dk}^{reg}(1 + \epsilon_k^{reg}))$$

$$\beta_{dv}^{res} \sim Gamma(\alpha_1 + z_{dv}^{res} + z_{dv}^{sum,res}, \alpha_2 + 1 + \epsilon_d^{res})$$

$$\epsilon_k^{reg} \sim Gamma(\alpha_5 + \sum_{d,v} \; z_{dv,k}^{sum,reg}, \alpha_6 + \sum_{d,v} \theta_{dk}^{reg} \beta_{kv}^{reg})$$

$$\epsilon_d^{res} \sim Gamma(\alpha_5 + \sum_v \; z_{dv}^{sum,res}, \alpha_6 + \sum_v \beta_{dv}^{res})$$

$$[\{z_{dv,k}^{reg}\}_k, z_{dv}^{res}] \sim Mult([\{\theta_{dk}^{reg} \beta_{kv}^{reg}\}_k, \beta_{dv}^{res}])$$

$$[\{z_{dv,k}^{sum,reg}\}_k, z_{dv}^{sum,res}] \sim Mult([\{\theta_{dk}^{reg} \beta_{kv}^{reg} \epsilon_k^{reg}\}_k, \beta_{dv}^{res} \epsilon_d^{res}])$$

(Note that in the benchmarks in which the offset variables are constrained to be constant, the posterior distribution of this parameter is: $\epsilon \sim Gamma(\alpha_5 + \sum_{d,v,k} \; z_{dv,k}^{sum,reg} + \sum_{d,v} z_{dv}^{sum,res}, \alpha_6 + \sum_{d,v,k} \theta_{dk}^{reg} \beta_{kv}^{reg} + \sum_{d,v} \beta_{dv}^{res})$.)

We estimate the model using variational inference, following Blei et al. (2003, 2016). The variational approximation of the topic assignment is also a multinomial distribution, with parameters $\phi_{dv}^{reg}, \phi_{dv}^{res}, \phi_{dv}^{sum,reg}, \phi_{dv}^{sum,res}$. The variational approximation of $\beta_v^{reg}, \beta_v^{res}$ are Gamma distributions, with parameters $\tilde{\beta}_v^{reg}, \tilde{\beta}_v^{res}$. The variational approximation of θ_d^{reg} are Gamma distributions, with parameters $\tilde{\theta}_d^{reg}$. The variational approximation of $\epsilon_k^{reg}, \epsilon_d^{res}$ are Gamma distributions, with parameters $\tilde{\epsilon}_k^{reg}, \tilde{\epsilon}_d^{res}$. The coordinate ascent mean-field variational inference algorithm updates these parameters as follows at each iteration (Blei et al., 2016):

$$\tilde{\theta}_{dk}^{reg} = < \alpha_3 + \textstyle\sum_v (w_{dv}\phi_{dv,k1}^{reg} + w_{dv}^{summary}\phi_{dv,k}^{summary,reg}), \alpha_4 + \sum_v \frac{\tilde{\beta}_{kv}^{regSHAPE}}{\tilde{\beta}_{kv}^{regRATE}}(1 + \frac{\tilde{\epsilon}_k^{regSHAPE}}{\tilde{\epsilon}_k^{regRATE}}) >$$

$$\tilde{\beta}_{kv}^{reg} = < \alpha_1 + \textstyle\sum_d (w_{dv}\phi_{dv,k}^{reg} + w_{dv}^{summary}\phi_{dv,k}^{summary,reg}), \alpha_2 + \sum_d \frac{\tilde{\theta}_{dk}^{regSHAPE}}{\tilde{\theta}_{dk}^{regRATE}}(1 + \frac{\tilde{\epsilon}_k^{regSHAPE}}{\tilde{\epsilon}_k^{regRATE}}) >$$

$$\tilde{\beta}_{dv}^{res} = < \alpha_1 + w_{dv}\phi_{dv}^{res} + w_{dv}^{summary}\phi_{dv}^{summary,res}, \alpha_2 + 1 + \frac{\tilde{\epsilon}_d^{resSHAPE}}{\tilde{\epsilon}_d^{resRATE}} >$$

$$\tilde{\epsilon}_k^{reg} = < \alpha_5 + \textstyle\sum_{d,v} (w_{dv}^{summary}\phi_{dv,k}^{summary,reg}), \alpha_6 + \sum_{d,v} \frac{\tilde{\theta}_{dk}^{regSHAPE}}{\tilde{\theta}_{dk}^{regRATE}} \frac{\tilde{\beta}_{kv}^{regSHAPE}}{\tilde{\beta}_{kv}^{regRATE}} >$$

$$\tilde{\epsilon}_d^{res} = < \alpha_5 + \textstyle\sum_v (w_{dv}^{summary}\phi_{dv}^{summary,res}), \alpha_6 + \sum_v \frac{\tilde{\beta}_{dv}^{resSHAPE}}{\tilde{\beta}_{dv}^{resRATE}} >$$

$$[\{\phi_{dv,k}^{reg}\}_k, \phi_{dv}^{res}] \propto exp(\{\Psi(\tilde{\theta}_{dk}^{regSHAPE}) - log(\tilde{\theta}_{dk}^{regRATE}) + \Psi(\tilde{\beta}_{kv}^{regSHAPE}) - log(\tilde{\beta}_{kv}^{regRATE})\}_k, \Psi(\tilde{\beta}_{dv}^{resSHAPE}) - log(\tilde{\beta}_{dv}^{resRATE}))$$

$$[\{\phi_{dv,k}^{sum,reg}\}_k, \phi_{dv}^{sum,res}] \propto exp(\{\Psi(\tilde{\theta}_{dk}^{regSHAPE}) - log(\tilde{\theta}_{dk}^{regRATE}) + \Psi(\tilde{\beta}_{kv}^{regSHAPE}) - log(\tilde{\beta}_{kv}^{regRATE}) + \Psi(\tilde{\epsilon}_k^{regSHAPE}) - log(\tilde{\epsilon}_k^{regRATE})\}_k, \Psi(\tilde{\beta}_{dv}^{resSHAPE}) - log(\tilde{\beta}_{dv}^{resRATE}) + \Psi(\tilde{\epsilon}_d^{resSHAPE}) - log(\tilde{\epsilon}_d^{resRATE}))$$

Where Ψ is the digamma function. Like Gopalan et al. (2014), we declare convergence when the change in likelihood is less than 0.001%. Note that this criterion is only used to assess convergence. The coordinate ascent mean-field variational inference algorithm maximizes Evidence Lower Bound (ELBO), which is equivalent to minimizing the KL divergence between the posterior distribution of the model parameters and the approximate distribution.

Following Gopalan et al. (2014), we set $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0.3$.

C Variational Inference: Predicting Summary Based on Full Text

We consider an out-of-sample document d^{out} for which the full text is available, but not the summary. The following auxiliary variables are introduced:

$z^{reg}_{d^{out}v,k} \sim Poisson(\theta^{reg}_{d^{out}k}\beta^{reg}_{kv})$; $z^{res}_{d^{out}v} \sim Poisson(\beta^{res}_{d^{out}v})$ such that $w_{d^{out}v} = \sum_k z^{reg}_{d^{out}v,k} + z^{res}_{d^{out}v}$.

The posterior distributions of the variables related to the out-of-sample document d^{out} are given as follows:

$\theta^{reg}_{d^{out}k} \sim Gamma(\alpha_3 + \sum_v z^{reg}_{d^{out}v,k}, \alpha_4 + \sum_v \beta^{reg}_{vk})$

$\beta^{res}_{d^{out}v} \sim Gamma(\alpha_1 + z^{res}_{d^{out}v}, \alpha_2 + 1)$

$[\{z^{reg}_{d^{out}v,k}\}_k, z^{res}_{d^{out}v}] \sim Mult([\{\theta^{reg}_{d^{out}k}\beta^{reg}_{kv}\}_k, \beta^{res}_{d^{out}v}])$

We estimate the model using variational inference. The coordinate ascent mean-field variational inference algorithm updates these parameters as follows at each iteration (Blei et al., 2016):

$\tilde{\theta}^{reg}_{d^{out}k} = \quad < \alpha_3 + \sum_v (w_{d^{out}v}\phi^{reg}_{d^{out}v,k}), \alpha_4 + \sum_v \frac{\tilde{\beta}^{reg\,SHAPE}_{kv}}{\tilde{\beta}^{reg\,RATE}_{kv}} >$

$\tilde{\beta}^{res}_{d^{out}v} = \quad < \alpha_1 + w_{d^{out}v}\phi^{res}_{d^{out}v}, \alpha_2 + 1 >$

$[\{\phi^{reg}_{d^{out}v,k}\}_k, \phi^{res}_{d^{out}v}] \propto exp(\{\Psi(\tilde{\theta}^{reg}_{d^{out}k}{}^{SHAPE}) - log(\tilde{\theta}^{reg}_{d^{out}k}{}^{RATE}) + \Psi(\tilde{\beta}^{reg\,SHAPE}_{kv}) - log(\tilde{\beta}^{reg\,RATE}_{kv})\}_k,$
$\Psi(\tilde{\beta}^{res}_{d^{out}v}{}^{SHAPE}) - log(\tilde{\beta}^{res}_{d^{out}v}{}^{RATE}))$

Like Gopalan et al. (2014), we declare convergence when the change in likelihood is less than 0.001%. Note that this criterion is only used to assess convergence. The coordinate

ascent mean-field variational inference algorithm maximizes Evidence Lower Bound (ELBO), which is equivalent to minimizing the KL divergence between the posterior distribution of the model parameters and the approximate distribution. Following Gopalan et al. (2014), we set $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0.3$.

D Latent Dirichlet Allocation

For each topic k, we define β_k as a $1 * L$ vector that we estimate, that contains the topic-word set of probability weights for topic k, that is, the probability that a token is assigned to each word given that it is assigned to topic k. For each document d, we define θ_d as a $1 * K$ vector that we estimate, that contains the document-topic set of probability weights for document d, that is, the probability that a token is assigned to each topic given that it is in document d. The i^{th} token in document d belongs to topic $z_i^d \in \{1, ...K\}$. The variable z_i^d is an unobserved, latent variable, which is also estimated. We denote by $w_i^d \in \{1, ...L\}$ the index of the word associated with the i^{th} token in document d. LDA does not include offset variables that capture variations in text between full documents and summaries. Accordingly, we assume that the full documents and their summaries have the topic intensities.

The data generating process assumed by LDA is as follows:

1. For each topic $k = 1, ...K$:

   - Choose $\beta_k \sim Dirichlet(\alpha_1)$

2. For each document $d = 1, ...D$:

- Choose $\theta_d \sim Dirichlet(\alpha_2)$
- For each token i in the main document or in the summary:
  - Select a topic $z_i^d \sim Multinomial(\theta_d)$
  - Select a word $\omega_i^d \sim Multinomial(\beta_{z_i^d})$

The priors on the topic-word probabilities $\{\phi_k\}$ and the document-topic probabilities $\{\theta_d\}$ are given as follows: $\beta_k \sim Dirichlet(\alpha_1)$; $\theta_d \sim Dirichlet(\alpha_2)$. In order to mimic the Poisson Factorization priors, we set α_1=0.3 and α_2=0.3. Given this specification, the posterior distributions of all variables are given in closed form:

$$Prob(z_i^d = k|\omega_i^d, \{\beta_k\}, \theta_d) = \frac{Prob(\omega_i^d|z_i^d = k, \beta_k)Prob(z_i^d = k|\theta_d)}{\sum_{k'} Prob(\omega_i^d|z_i^d = k', \beta_{k'})Prob(z_i^d = k'|\theta_d)} = \frac{\beta_{k\omega_i^d}\theta_{dk}}{\sum_{k'} \beta_{k'\omega_i^d}\theta_{dk'}} \quad \text{(WA1)}$$

$$Prob(\beta_k|\{z_i^d\}, \{\omega_i^d\}) = Dirichlet(\alpha_1 + \sum_{(i,d):z_i^d=k} 1(\omega_i^d = 1),\ \ldots\ , \alpha_1 + \sum_{(i,d):z_i^d=k} 1(\omega_i^d = L)) \quad \text{(WA2)}$$

$$Prob(\theta_d|\{z_i^d\}) = Dirichlet(\alpha_2 + \sum_i 1(z_i^d = 1), \ldots \alpha_2 + \sum_i 1(z_i^d = K)) \quad \text{(WA3)}$$

We estimate the model using variational inference, following Blei et al. (2003, 2016). The variational approximation of the topic assignment is also a multinomial distribution, with parameters ϕ_{dv}. The variational approximation of the topic distribution is a Dirichlet, with parameter $\tilde{\beta}_v$. The variational approximation of the topic proportions is also a Dirichlet, with parameter $\tilde{\theta}_d$. The coordinate ascent mean-field variational inference algorithm updates these parameters as follows at each iteration, until convergence (Blei et al., 2003, 2016):

$$\{\phi_{dv,k}\}_k \propto\ exp(\Psi(\tilde{\beta}_{kv}) - \Psi(\textstyle\sum_v \tilde{\beta}_{kv}) + \Psi(\tilde{\theta}_{dk}))$$

$$\tilde{\beta}_{kv} = \quad \alpha_1 + \textstyle\sum_d \phi_{dv,k}(w_{dv} + w_{dv}^{summary})$$

$$\tilde{\theta}_{dk} = \quad \alpha_2 + \sum_v \phi_{dv,k}(w_{dv} + w_{dv}^{summary})$$

where Ψ is the digamma function and w_{dv} and $w_{dv}^{summary}$ are defined as in Poisson Factorization as the number of occurrences of word v in document d. Like with the proposed model, we set the number of topics to 100.

We test performance using perplexity as we do for Poisson Factorization.

$$Perplexity = exp(-\frac{\sum_{d \in D_{test}} \sum_{obs \in d} log(\sum_k \varphi_{k,obs} \hat{\theta}_{dk})}{N}) \tag{WA4}$$

Where $\sum_k \varphi_{k,obs} \hat{\theta}_{dk}$ is the probability of seeing word *obs* in document d, based on the estimated topic intensities $\hat{\theta}_d$.

E Sensitivity to Prior Parameters

Throughout the paper, we following Gopalan et al. (2014) and set $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0.3$.

E.1 More Diffuse Prior

We test a prior that is even more diffuse: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0.2$ (the variance of the prior is 5 rather than 3.33). The prior being more diffuse, there is less penalty for introducing "non-flat" topics, and as a result we find more non-flat regular topics with this prior. In the main model, we find 60 non-flat regular topics in the marketing academic papers datasets, 52 in the movies dataset, and 48 in the TV shows dataset. Table WA4 reports the fit and predictive performance of the various benchmarks under this alternative prior. Results are qualitatively similar to those in Table 2.

Table WA4: Fit and Predictive Performance - Marketing Academic Papers. More Diffuse Prior.

		Fit			Predictive perf.
	Approach	**Calibration documents**	**Calibration summaries**	**Validation documents**	**Validation summaries**
Marketing academic papers	Full Model	103.43	64.13	107.20	80.27
	No residual topic	209.28	154.23	237.62	180.18
	ϵ constant	103.45	71.36	107.17	82.84
	No residual topic and ϵ constant	208.69	156.81	237.72	185.58
	Residual topics only	98.89	68.29	103.38	81.98
	LDA	211.06	158.71	237.73	185.40
Movies	Full Model	170.35	172.53	176.06	311.77
	No residual topic	280.43	288.39	332.34	366.15
	ϵ constant	170.61	215.24	176.14	339.92
	No residual topic and ϵ constant	279.35	365.93	333.30	439.91
	Residual topics only	158.14	197.95	165.67	361.86
	LDA	279.26	360.53	332.75	433.21
TV show episodes	Full Model	244.21	240.93	239.18	400.43
	No residual topic	377.19	449.78	369.99	483.34
	ϵ constant	244.66	310.42	239.26	571.48
	No residual topic and ϵ constant	376.09	690.34	369.03	703.78
	Residual topics only	221.21	276.10	219.94	598.60
	LDA	376.33	690.99	368.69	702.19

Note: fit and predictive performance are measured using perplexity (lower values indicate better fit).

E.2 Less Diffuse Prior

We replicate our analysis with a prior that is less diffuse: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0.4$ (the variance of the prior is 2.5 rather than 3.33). With a less diffuse prior, there is more penalty for including non-flat regular topics. As a result, in the main model, we find 15 non-flat regular topics in the marketing academic papers datasets, 15 in the movies dataset, and 17 in the TV shows dataset. Table WA5 reports the fit and predictive performance of the various benchmarks under this alternative prior. Results are qualitatively similar to those in Table 2.

Table WA5: Fit and Predictive Performance - Marketing Academic Papers. Less Diffuse Prior.

		Fit			Predictive perf.
	Approach	**Calibration documents**	**Calibration summaries**	**Validation documents**	**Validation summaries**
Marketing academic papers	Full Model	107.90	70.09	113.37	83.53
	No residual topic	195.59	140.21	226.20	169.93
	ϵ constant	107.92	76.12	113.60	88.93
	No residual topic and ϵ constant	195.69	146.09	225.67	176.90
	Residual topics only	104.72	73.32	109.93	86.75
	LDA	194.95	144.11	225.16	175.02
Movies	Full Model	174.73	186.25	186.28	298.00
	No residual topic	262.75	271.29	320.19	362.00
	ϵ constant	175.08	221.56	186.19	350.70
	No residual topic and ϵ constant	261.22	339.87	319.38	423.87
	Residual topics only	169.34	211.54	178.90	353.55
	LDA	262.06	338.76	320.21	420.55
TV show episodes	Full Model	251.08	257.26	250.50	424.95
	No residual topic	357.06	417.25	356.73	438.10
	ϵ constant	251.37	329.93	250.44	578.29
	No residual topic and ϵ constant	357.37	629.47	356.82	689.35
	Residual topics only	247.67	311.14	246.56	564.98
	LDA	349.72	609.93	349.66	683.18

Note: fit and predictive performance are measured using perplexity (lower values indicate better fit).

F Alternative Version without Offset Variables

In order to further explore the role of offset variables, we estimate an alternative model in which full documents and summaries are treated as independent documents. That is, instead of modeling the content of D documents and their associated summaries, this alternative model treats the D documents and the D summaries as independent documents. The data generating process is as follows:

1. For each regular topic $k = 1, ...K$:

- For each word v, draw $\beta_{kv}^{reg} \sim Gamma(\alpha_1, \alpha_2)$

2. For each document $d = 1, ...D$:

- For each regular topic, draw topic intensity in full document: $\theta_{dk}^{reg} \sim Gamma(\alpha_3, \alpha_4)$
- For each word v, draw residual topic weight in full document: $\beta_{dv}^{res} \sim Gamma(\alpha_1, \alpha_2)$
- For each word v, draw word count in full document: $w_{dv} \sim Poisson(\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} + \beta_{dv}^{res})$
- For each regular topic, draw topic intensity in summary: $\theta_{dk}^{sum,reg} \sim Gamma(\alpha_3, \alpha_4)$
- For each word v, draw residual topic weight in summary: $\beta_{dv}^{sum,res} \sim Gamma(\alpha_1, \alpha_2)$
- For each word v, draw word count in summary: $w_{dv}^{sum} \sim Poisson(\sum_k \theta_{dk}^{sum,reg} \beta_{kv}^{reg} + \beta_{dv}^{sum,res})$

We estimate this alternative model on all three datasets. As this model does not allow predicting the content of a summary based on the content of a full document, we cannot compare its performance to the main model. Tables WA6 to WA8 are the equivalent of Tables WA1 to WA3 for the main model, i.e., they report words with the largest weights for each non-flat regular topic. As there is no offset variable in this alternative specification, we compute for each regular topic the average across documents of the ratio of the topic intensity in the summary to the topic intensity in the full document: $\frac{1}{D} \sum_d \frac{\theta_{dk}^{sum,reg}}{\theta_{dk}^{reg}}$. We sort topics according to this metric. We see that the topics learned from this alternative specification are quite different from the topics learned from the main specification. This suggests that the presence of offset variables in the model has some influence on the topics

learned, as topics are defined as groups of words that not only tend to appear together, but that also tend to appear with the same relative frequency in summaries vs. full documents.

Table WA6: Non-Flat Regular Topics - Marketing Academic Papers. Alternative Version Without Offset Variables.

Av. $\frac{\textbf{topic intensity in summary}}{\textbf{topic intensity in document}}$	**Av. topic intensity**	**Words with largest weights**
0.944	0.500	consumer
0.863	0.502	consumers, positive, negative
0.764	0.505	product, products, participants, design, consumer
0.731	0.505	model, models, value, values, function
0.706	0.507	product, products, conditions, condition
0.690	0.517	firms, market, value, model, sales
0.676	0.538	marketing, data, variables, model, firm
0.672	0.511	brand, brands, consumer, table, marketing
0.669	0.532	data, model, table, average, estimates
0.666	0.518	price, prices, model, market, consumer
0.660	0.528	product, products, data, model, table
0.657	0.529	model, data, models, parameters, table
0.649	0.516	positive, behavior, interaction, negative, measures
0.610	0.512	consumers, consumer, low, conditions, value
0.506	0.555	product, consumer, products, model, purchase
0.466	0.611	participants, condition, people

Note: topics are sorted in decreasing order of the average ratio of the topic intensity in summaries to the topic intensity in documents, which mimics the offset variable. The second column reports the average intensity of that topic across documents (i.e., full texts and summaries). Only words with weights β_{kv}^{reg} larger than 1% of the sum of the weights for that topic ($\sum_v \beta_{kv}^{reg}$), are included in the table.

Table WA7: Non-Flat Regular Topics - Movies. Alternative Version Without Offset Variables.

Av. $\frac{\text{topic intensity in summary}}{\text{topic intensity in document}}$	Av. topic intensity	Words with largest weights
0.742	0.258	street, gun, begins, blood, beside
0.738	0.250	gonna, hey, night, shit, house
0.701	0.248	gonna, cut, shit, em, money
0.690	0.250	hey, guy, crowd, cut, house
0.677	0.262	night, begins, others, blood, rises
0.671	0.268	night, begins, hey, spins, swings
0.602	0.247	beat, gonna, hey, shit, guy
0.585	0.337	gun, police, street, shit, cars
0.512	0.326	street, night, apartment, hey, guy
0.465	0.338	night, street, house, bedroom, cut
0.426	0.280	cut, night, dissolve, begins, doors

Note: topics are sorted in decreasing order of the average ratio of the topic intensity in summaries to the topic intensity in documents, which mimics the offset variable. The second column reports the average intensity of that topic across documents (i.e., full texts and summaries). Only words with weights β_{kv}^{reg} larger than 1% of the sum of the weights for that topic ($\sum_v \beta_{kv}^{reg}$), are included in the table.

Table WA8: Non-Flat Regular Topics - TV Shows. Alternative Version Without Offset Variables.

Av. $\frac{\text{topic intensity in summary}}{\text{topic intensity in document}}$	Av. topic intensity	Words with largest weights
0.898	0.975	gonna
0.737	0.975	okay
0.627	0.976	like, know, want, gonna, really
0.578	0.975	know, like, really, would, want
0.554	0.975	yeah, like, know, hey, want
0.539	0.975	okay, gonna, yeah, know, hey
0.509	0.975	know, like, yeah, want, hey
0.502	0.976	like, gonna, know, yeah, really

Note: topics are sorted in decreasing order of the average ratio of the topic intensity in summaries to the topic intensity in documents, which mimics the offset variable. The second column reports the average intensity of that topic across documents (i.e., full texts and summaries). Only words with weights β_{kv}^{reg} larger than 1% of the sum of the weights for that topic ($\sum_v \beta_{kv}^{reg}$), are included in the table.

We also explore the link between the various distinctiveness measures computed from this

alternative specification, and success in each dataset. "Inside the cone distinctiveness" and "outside the cone distinctiveness" are defined similarly as with the main model, based on the topic intensities estimated for each full document. "Outside the cone emphasis in summary" is not defined for this alternative model, and hence it is not included in the analysis. In the marketing academic papers dataset, we find that the correlation in distinctiveness measures between the original model and this alternative model is 0.87 for "inside the cone distinctiveness" ($p < 0.01$), 0.80 for "outside the cone distinctiveness" ($p < 0.01$). For movies, these correlation are respectively 0.56 ($p < 0.01$), and 0.83 ($p < 0.01$); for TV shows they are 0.62 ($p < 0.01$) and 0.93 ($p < 0.01$). Results are reported in Tables WA9 to WA11.

Table WA9: Link Between Distinctiveness Measures and Citations - Marketing Academic Papers. Alternative Version Without Offset Variables.

Covariates	**DV=log(1+#citations)**
Journal fixed effects	✓
Publication year fixed effects	✓
Intensities of paper on non-flat regular topics	✓
Number of pages in paper	0.049**
"Inside the cone distinctiveness" (from journal)	-0.010
"Outside the cone distinctiveness"	0.151**
Number of parameters	28
Number of observations	1,000
R^2	0.313

Note: OLS regression. *: significant at $p < 0.10$. **: significant at $p < 0.05$. Both distinctiveness measures are standardized across papers for interpretability.

Table WA10: Link Between Distinctiveness Measures and Performance - Movies. Alternative version Without Offset Variables.

Covariates	**DV=movie rating**	**DV=log(return on investment)**
MPAA rating fixed effects	✓	✓
Genre fixed effects	✓	✓
Intensities of script on non-flat regular topics	✓	✓
Movie duration (in min)	0.002	-0.002
Log(inflation-adjusted production budget)	-0.057	-0.303**
Movie rating	–	0.457**
"Inside the cone distinctiveness" (from genre)	-0.069	-0.152**
"Outside the cone distinctiveness"	0.223**	-0.207
Number of parameters	40	41
Number of observations	596	581
R^2	0.308	0.241

Note: OLS regression. *: significant at $p < 0.10$. **: significant at $p < 0.05$. Both distinctiveness measures and movie ratings are standardized across movies for interpretability.

Table WA11: Link Between Distinctiveness Measures and Performance - TV Episodes. Alternative Version Without Offset Variables.

Covariates	**DV=episode rating**
TV series fixed effects	✓
Intensities of script on non-flat regular topics	✓
"Inside the cone distinctiveness" (from TV series)	-0.011
"Outside the cone distinctiveness"	0.070**
Number of parameters	328
Number of observations	9,285
R^2	0.686

Note: OLS regression. *: significant at $p < 0.10$. **: significant at $p < 0.05$. Both distinctiveness measures and episode ratings are standardized across episodes for interpretability.

G Dynamic Version of the Model

G.1 Data Generating Process

We assume that documents have a discrete time index $t = 1, ...T$. We introduce dynamics in a manner inspired by Blei and Lafferty (2006). That is, we model the weights of a topic across words in each time period as equal to the weights in the previous period, plus a set of additional weights. We achieve this by introducing an additional set of offset variables, that quantify how a topic evolves over time. That is, the weight of topic k on word v, instead of being constant across time periods and equal to β_{kv}^{reg}, now evolves over time, and is equal to: $\beta_{kv1}^{reg} + \sum_{\tau=1}^{t} \eta_{kv\tau}^{reg}$ in period t. The data generating process is updated as follows:

1. For each regular topic $k = 1, ...K$:
   - For each word v, draw base topic weight: $\beta_{kv1}^{reg} \sim Gamma(\alpha_1, \alpha_2)$
   - For each word v and each time period $t > 1$, draw time-specific offset weight: $\eta_{kv\tau}^{reg} \sim Gamma(\alpha_1, \alpha_2)$
   - Draw offset variable $\epsilon_k^{reg} \sim Gamma(\alpha_5, \alpha_6)$
2. For each residual topic $d = 1, ...D$:
   - For each word v, draw $\beta_{dv}^{res} \sim Gamma(\alpha_1, \alpha_2)$
   - Draw offset variable $\epsilon_d^{res} \sim Gamma(\alpha_5, \alpha_6)$
3. For each document $d = 1, ...D$ in time period t_d:
   - For each regular topic, draw topic intensity $\theta_{dk}^{reg} \sim Gamma(\alpha_3, \alpha_4)$

- For each word v, draw word count $w_{dv} \sim Poisson(\sum_k \theta_{dk}^{reg}(\beta_{kv1}^{reg} + \sum_{\tau=2}^{t_d} \eta_{kv\tau}^{reg}) + \beta_{dv}^{res})$

4. For each document summary $d = 1, ...D$:

- For each word v, draw word count $w_{dv}^{summary} \sim Poisson(\sum_k \theta_{dk}^{reg}(\beta_{kv1}^{reg} + \sum_{\tau=2}^{t_d} \eta_{kv\tau}^{reg})\epsilon_k^{reg} + \beta_{dv}^{res}\epsilon_d^{res})$

G.2 Variational Inference for Model Estimation

In order to estimate the model, we start by defining auxiliary variables that allocate the occurrences of each word v in each document d across the various topics: $z_{dv,k1}^{reg} \sim Poisson(\theta_{dk}^{reg}\beta_{kv1}^{reg})$;$z_{dv,k\tau}^{reg} \sim Poisson(\theta_{dk}^{reg}\eta_{kv\tau}^{reg})$; $z_{dv}^{res} \sim Poisson(\beta_{dv}^{res})$, such that $w_{dv} = \sum_k(z_{dv,k1}^{reg} + \Sigma_{\tau=2}^{t_d} z_{dv,k\tau}^{reg}) + z_{dv}^{res}$. Similar variables are defined for the summaries: $z_{dv,k1}^{sum,reg} \sim Poisson(\theta_{dk}^{reg}\beta_{kv1}^{reg}\epsilon_k^{reg})$; $z_{dv,k\tau}^{sum,reg} \sim Poisson(\theta_{dk}^{reg}\eta_{kv\tau}^{reg}\epsilon_k^{reg})$; $z_{dv}^{sum,res} \sim Poisson(\beta_{dv}^{res}\epsilon_d^{res})$, such that $w_{dv}^{summary} = \sum_k(z_{dv,k}^{sum,reg} + \Sigma_{\tau=2}^{t_d} z_{dv,k\tau}^{sum,reg}) + z_{dv}^{sum,res}$. With the addition of these auxiliary variables, the model is conditionally conjugate, i.e., the posterior distribution of each parameter conditional on the other parameters and the data, is given in closed form. The conditional posterior distributions are given as follows (based on Gopalan et al., 2013, 2014):

$$\theta_{dk}^{reg} \sim Gamma(\alpha_3 + \sum_v(z_{dv,k1}^{reg} + \Sigma_{\tau=2}^{t_d} z_{dv,k\tau}^{reg} + z_{dv,k1}^{sum,reg} + \Sigma_{\tau=2}^{t_d} z_{dv,k\tau}^{sum,reg}), \alpha_4 + \sum_v(\beta_{kv1}^{reg} + \Sigma_{\tau=2}^{t_d}\eta_{kv\tau}^{reg})(1 + \epsilon_k^{reg}))$$

$$\beta_{kv1}^{reg} \sim Gamma(\alpha_1 + \sum_d(z_{dv,k1}^{reg} + z_{dv,k1}^{sum,reg}), \alpha_2 + \sum_d \theta_{dk}^{reg}(1 + \epsilon_k^{reg}))$$

$$\eta_{kv\tau}^{reg} \sim Gamma(\alpha_7 + \sum_{d:t_d \geq \tau}(z_{dv,k\tau}^{reg} + z_{dv,k\tau}^{sum,reg}), \alpha_8 + \sum_{d:t_d \geq \tau} \theta_{dk}^{reg}(1 + \epsilon_k^{reg}))$$

$$\beta_{dv}^{res} \sim Gamma(\alpha_1 + z_{dv}^{res} + z_{dv}^{sum,res}, \alpha_2 + 1 + \epsilon_d^{res})$$

$$\epsilon_k^{reg} \sim Gamma(\alpha_5 + \sum_{d,v}(z_{dv,k1}^{sum,reg} + \Sigma_{\tau=2}^{t_d} z_{dv,k\tau}^{sum,reg}), \alpha_6 + \sum_{d,v} \theta_{dk}^{reg}(\beta_{kv1}^{reg} + \Sigma_{\tau=2}^{t_d}\eta_{kv\tau}^{reg}))$$

$$\epsilon_d^{res} \sim Gamma(\alpha_5 + \sum_v z_{dv}^{sum,res}, \alpha_6 + \sum_v \beta_{dv}^{res})$$

$$[\{z_{dv,k1}^{reg}\}_k, \{z_{dv,k\tau}^{reg}\}_{k,\tau}, z_{dv}^{res}] \sim Mult([\{\theta_{dk}^{reg}\beta_{kv1}^{reg}\}_k, \{\theta_{dk}^{reg}\eta_{kv\tau}^{reg}\}_{k,\tau}, \beta_{dv}^{res}])$$

$$[\{z_{dv,k1}^{sum,reg}\}_k, \{z_{dv,k\tau}^{sum,reg}\}_{k,\tau}, z_{dv}^{sum,res}] \sim Mult([\{\theta_{dk}^{reg}\beta_{kv1}^{reg}\epsilon_k^{reg}\}_k, \{\theta_{dk}^{reg}\eta_{kv\tau}^{reg}\epsilon_k^{reg}\}_{k,\tau}, \beta_{dv}^{res}\epsilon_d^{res}])$$

We estimate the model using variational inference, following Blei et al. (2003, 2016). The variational approximation of the topic assignment is also a multinomial distribution, with parameters $\phi_{dv}^{reg}, \phi_{dv}^{res}, \phi_{dv}^{sum,reg}, \phi_{dv}^{sum,res}$. The variational approximation of $\beta_v^{reg}, \beta_v^{res}$ are Gamma distributions, with parameters $\tilde{\beta}_v^{reg}, \tilde{\beta}_v^{res}$. The variational approximation of η_v^{reg} are Gamma distributions, with parameters $\tilde{\eta}_v^{reg}$. The variational approximation of θ_d^{reg} are Gamma distributions, with parameters $\tilde{\theta}_d^{reg}$. The variational approximation of $\epsilon_k^{reg}, \epsilon_d^{res}$ are Gamma distributions, with parameters $\tilde{\epsilon}_k^{reg}, \tilde{\epsilon}_d^{res}$. The coordinate ascent mean-field variational inference algorithm updates these parameters as follows at each iteration (Blei et al., 2016):

$$\tilde{\theta}_{dk}^{reg} = < \alpha_3 + \sum_v (w_{dv}(\phi_{dv,k1}^{reg} + \Sigma_{\tau=2}^{t_d}\phi_{dv,k\tau}^{reg}) + w_{dv}^{summary}(\phi_{dv,k1}^{summary,reg} + \Sigma_{\tau=2}^{t_d}\phi_{dv,k\tau}^{summary,reg})), \alpha_4 + \sum_v (\frac{\tilde{\beta}_{kv1}^{reg\,SHAPE}}{\tilde{\beta}_{kv1}^{reg\,RATE}} + \Sigma_{\tau=2}^{t_d}\frac{\tilde{\eta}_{kv\tau}^{reg\,SHAPE}}{\tilde{\eta}_{kv\tau}^{reg\,RATE}})(1 + \frac{\tilde{\epsilon}_k^{reg\,SHAPE}}{\tilde{\epsilon}_k^{reg\,RATE}}) >$$

$$\tilde{\beta}_{kv1}^{reg} = < \alpha_1 + \sum_d (w_{dv}\phi_{dv,k1}^{reg} + w_{dv}^{summary}\phi_{dv,k1}^{summary,reg}), \alpha_2 + \sum_d \frac{\tilde{\theta}_{dk}^{reg\,SHAPE}}{\tilde{\theta}_{dk}^{reg\,RATE}}(1 + \frac{\tilde{\epsilon}_k^{reg\,SHAPE}}{\tilde{\epsilon}_k^{reg\,RATE}}) >$$

$$\tilde{\eta}_{kv\tau}^{reg} = < \alpha_7 + \sum_{d:t_d \geq \tau} (w_{dv}\phi_{dv,k\tau}^{reg} + w_{dv}^{summary}\phi_{dv,k\tau}^{summary,reg}), \alpha_8 + \sum_{d:t_d \geq \tau} \frac{\tilde{\theta}_{dk}^{reg\,SHAPE}}{\tilde{\theta}_{dk}^{reg\,RATE}}(1 + \frac{\tilde{\epsilon}_k^{reg\,SHAPE}}{\tilde{\epsilon}_k^{reg\,RATE}}) >$$

$$\tilde{\beta}_{dv}^{res} = < \alpha_1 + w_{dv}\phi_{dv}^{res} + w_{dv}^{summary}\phi_{dv}^{summary,res}, \alpha_2 + 1 + \frac{\tilde{\epsilon}_d^{res\,SHAPE}}{\tilde{\epsilon}_d^{res\,RATE}} >$$

$$\tilde{\epsilon}_k^{reg} = < \alpha_5 + \sum_{d,v} w_{dv}^{summary}(\phi_{dv,k1}^{summary,reg} + \Sigma_{\tau=2}^{t_d}\phi_{dv,k\tau}^{summary,reg}), \alpha_6 + \sum_{d,v} \frac{\tilde{\theta}_{dk}^{reg\,SHAPE}}{\tilde{\theta}_{dk}^{reg\,RATE}}(\frac{\tilde{\beta}_{kv1}^{reg\,SHAPE}}{\tilde{\beta}_{kv1}^{reg\,RATE}} + \Sigma_{\tau=2}^{t_d}\frac{\tilde{\eta}_{kv\tau}^{reg\,SHAPE}}{\tilde{\eta}_{kv\tau}^{reg\,RATE}}) >$$

$$\tilde{\epsilon}_d^{res} = < \alpha_5 + \sum_v (w_{dv}^{summary}\phi_{dv}^{summary,res}), \alpha_6 + \sum_v \frac{\tilde{\beta}_{dv}^{res\,SHAPE}}{\tilde{\beta}_{dv}^{res\,RATE}} >$$

$$[\{\phi_{dv,k1}^{reg}\}_k, \{\phi_{dv,k\tau}^{reg}\}_{k,\tau}, \phi_{dv}^{res}] \propto exp(\{\Psi(\tilde{\theta}_{dk}^{reg\,SHAPE}) - log(\tilde{\theta}_{dk}^{reg\,RATE}) + \Psi(\tilde{\beta}_{kv1}^{reg\,SHAPE}) -$$

$log(\tilde{\beta}_{kv1}^{reg\,RATE})\}_k, \{\Psi(\tilde{\theta}_{dk}^{reg\,SHAPE}) - log(\tilde{\theta}_{dk}^{reg\,RATE}) + \Psi(\tilde{\eta}_{kv\tau}^{reg\,SHAPE}) - log(\tilde{\eta}_{kv\tau}^{reg\,RATE})\}_{k,\tau}$,

$\Psi(\tilde{\beta}_{dv}^{res\,SHAPE}) - log(\tilde{\beta}_{dv}^{res\,RATE}))$

$$[\{\phi_{dv,k1}^{sum,reg}\}_k, \{\phi_{dv,k\tau}^{sum,reg}\}_{k,\tau}, \phi_{dv}^{sum,res}] \propto exp(\{\Psi(\tilde{\theta}_{dk}^{reg\,SHAPE}) - log(\tilde{\theta}_{dk}^{reg\,RATE}) + \Psi(\tilde{\beta}_{kv1}^{reg\,SHAPE}) - log(\tilde{\beta}_{kv1}^{reg\,RATE}) + \Psi(\tilde{\epsilon}_k^{reg\,SHAPE}) - log(\tilde{\epsilon}_k^{reg\,RATE})\}_k, \{\Psi(\tilde{\theta}_{dk}^{reg\,SHAPE}) - log(\tilde{\theta}_{dk}^{reg\,RATE}) + \Psi(\tilde{\eta}_{kv\tau}^{reg\,SHAPE}) - log(\tilde{\eta}_{kv\tau}^{reg\,RATE}) + \Psi(\tilde{\epsilon}_k^{reg\,SHAPE}) - log(\tilde{\epsilon}_k^{reg\,RATE})\}_{k,\tau}, \Psi(\tilde{\beta}_{dv}^{res\,SHAPE}) - log(\tilde{\beta}_{dv}^{res\,RATE}) + \Psi(\tilde{\epsilon}_d^{res\,SHAPE}) - log(\tilde{\epsilon}_d^{res\,RATE}))$$

Where Ψ is the digamma function. Like Gopalan et al. (2014), we declare convergence when the change in likelihood is less than 0.001%. Note that this criterion is only used to assess convergence. Following Gopalan et al. (2014), we set $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = \alpha_7 = \alpha_8 = 0.3$.

Finally, we consider an out-of-sample document d^{out} for which the full text is available, but not the summary. The following auxiliary variables are introduced:

$z_{d^{out}v,k1}^{reg} \sim Poisson(\theta_{d^{out}k}^{reg}\beta_{kv1}^{reg})$;$z_{d^{out}v,k\tau}^{reg} \sim Poisson(\theta_{d^{out}k}^{reg}\eta_{kv\tau}^{reg})$; $z_{d^{out}v}^{res} \sim Poisson(\beta_{d^{out}v}^{res})$ such that $w_{d^{out}v} = \sum_k (z_{d^{out}v,k}^{reg} + \Sigma_{\tau=2}^{t_{d^{out}}} z_{d^{out}v,k\tau}^{reg}) + z_{d^{out}v}^{res}$.

The posterior distributions of the variables related to the out-of-sample document d^{out} are given as follows:

$$\theta_{d^{out}k}^{reg} \sim Gamma(\alpha_3 + \sum_v (z_{d^{out}v,k1}^{reg} + \Sigma_{\tau=2}^{t_{d^{out}}} z_{d^{out}v,k\tau}^{reg}), \alpha_4 + \sum_v (\beta_{vk1}^{reg} + \Sigma_{\tau=2}^{t_{d^{out}}} \eta_{kv\tau}^{reg}))$$

$$\beta_{d^{out}v}^{res} \sim Gamma(\alpha_1 + z_{d^{out}v}^{res}, \alpha_2 + 1)$$

$$[\{z_{d^{out}v,k1}^{reg}\}_k, \{z_{d^{out}v,k\tau}^{reg}\}_{k,\tau}, z_{d^{out}v}^{res}] \sim Mult([\{\theta_{d^{out}k}^{reg}\beta_{kv1}^{reg}\}_k, \{\theta_{d^{out}k}^{reg}\eta_{kv\tau}^{reg}\}_{k,\tau}, \beta_{d^{out}v}^{res}])$$

We estimate these parameters using variational inference. The coordinate ascent mean-field variational inference algorithm updates these parameters as follows at each iteration (Blei et al., 2016):

$$\tilde{\theta}^{reg}_{d^{out}k} = \ < \alpha_3 + \sum_v (w_{d^{out}v}(\phi^{reg}_{d^{out}v,k1} + \Sigma^{t_{d^{out}}}_{\tau=2} \phi^{reg}_{d^{out}v,k\tau})), \alpha_4 + \sum_v (\frac{\tilde{\beta}^{reg\,SHAPE}_{kv}}{\tilde{\beta}^{reg\,RATE}_{kv}} + \Sigma^{t_{d^{out}}}_{\tau=2} \frac{\tilde{\eta}^{reg\,SHAPE}_{kv\tau}}{\tilde{\eta}^{reg\,RATE}_{kv\tau}}) >$$

$$\tilde{\beta}^{res}_{d^{out}v} = \ < \alpha_1 + w_{d^{out}v}\phi^{res}_{d^{out}v}, \alpha_2 + 1 >$$

$$[\{\phi^{reg}_{d^{out}v,k1}\}_k, \{\phi^{reg}_{d^{out}v,k\tau}\}_{k,\tau}, \phi^{res}_{d^{out}v}] \propto exp(\{\Psi(\tilde{\theta}^{reg}_{d^{out}k}{}^{SHAPE}) - log(\tilde{\theta}^{reg}_{d^{out}k}{}^{RATE}) + \Psi(\tilde{\beta}^{reg\,SHAPE}_{kv1}) - log(\tilde{\beta}^{reg\,RATE}_{kv1})\}_k, \{\Psi(\tilde{\theta}^{reg}_{d^{out}k}{}^{SHAPE}) - log(\tilde{\theta}^{reg}_{d^{out}k}{}^{RATE}) + \Psi(\tilde{\eta}^{reg\,SHAPE}_{kv\tau}) - log(\tilde{\eta}^{reg\,RATE}_{kv\tau})\}_{k,\tau},$$
$$\Psi(\tilde{\beta}^{res}_{d^{out}v}{}^{SHAPE}) - log(\tilde{\beta}^{res}_{d^{out}v}{}^{RATE}))$$

G.3 Results

We apply this extension of the model to the Marketing academic papers dataset, in which we have a census of all papers published in a 5-year window in four journals. (In the movie dataset, we only have a subset of the movies produced each year, which would limit our ability to study dynamics in topics - in the TV shows dataset we do not have the year in which each episode was produced.)

Table WA12 is similar to Table 2, for the dynamic model. Note that LDA does not capture dynamics, hence it is the same as in Table 2. The "residual topics only" version is also unchanged. Similar to the case without dynamics, we find that the proposed model performs best in terms of fitting the summaries of calibration documents and predicting the summaries of validation documents. We again find that the introduction of residual topics greatly improves fit and predictive performance, and that "residual topics only" performs best in terms of fitting the full documents, with the same limitations as in the non-dynamic case (no topic is learned across topic, and poorer performance at fitting or predicting the content of summaries).

Table WA12: Fit and Predictive Performance - Marketing Academic Papers.

	Fit			**Predictive perf.**
Approach	**Calibration documents**	**Calibration summaries**	**Validation documents**	**Validation summaries**
Full Model	104.50	69.94	109.69	84.10
No residual topic	151.44	110.33	295.73	243.61
ϵ constant	104.48	73.18	109.75	86.14
No residual topic and ϵ constant	151.03	111.36	296.53	246.97
Residual topics only	101.83	70.84	106.68	84.30
LDA	197.13	145.73	227.12	176.34

Note: fit and predictive performance are measured using perplexity (lower values indicate better fit).

Table WA13 is the equivalent of Table 4, for the dynamic version of the model. "Inside the cone distinctiveness" and "Outside the cone emphasis in summary" are defined as in the main model without dynamics. "Outside the cone distinctiveness" is defined similarly as well: $\frac{\sum_v \beta_{dv}^{res}}{\sum_v [\sum_k \theta_{dk}^{reg} (\beta_{kv1}^{reg} + \sum_{\tau=2}^{t_d} \eta_{kv\tau}^{reg}) + \beta_{dv}^{res}]}$. Results are similar to the ones obtained with the non-dynamic version of the model: "inside the cone distinctiveness," " outside the cone distinctiveness" and "outside the cone emphasis in summary" are all positively and significantly related to the number of citations.

Table WA13: Link Between Distinctiveness Measures and Citations - Marketing Academic Papers. Dynamic Version of the Model.

Covariates	**DV=log(1+#citations)**
Journal fixed effects	✓
Publication year fixed effects	✓
Intensities of paper on non-flat regular topics	✓
Number of pages in paper	0.048**
"Inside the cone distinctiveness" (from journal)	0.073**
"Outside the cone distinctiveness"	0.126**
"Outside the cone emphasis in summary"	0.063**
Number of parameters	19
Number of observations	1,000
R^2	0.290

Note: OLS regression. *: significant at $p < 0.10$. **: significant at $p < 0.05$. All three distinctiveness measures are standardized across papers for interpretability.

H Alternative Vocabulary Sizes, Measures and Analyses

H.1 Alternative Number of Words in Vocabulary

H.1.1 2,000 Words in Vocabulary

We increase the number of words in the vocabulary to 2,000, keeping the same *tf-idf* approach for selecting words. Table WA14 reports the fit and predictive performance of the various benchmarks under this alternative vocabulary size. Results are qualitatively similar to those in Table 2.

In the main model, we find 24 non-flat regular topics in the marketing academic papers datasets, 18 in the movies dataset, and 27 in the TV shows dataset. Tables WA15 to WA17 replicate Tables 4 to 6 with 2,000 words in the vocabulary. The correlation in the distinc-

Table WA14: Fit and Predictive Performance - Marketing Academic Papers. 2,000 Words in Vocabulary.

		Fit			Predictive perf.
	Approach	**Calibration documents**	**Calibration summaries**	**Validation documents**	**Validation summaries**
Marketing academic papers	Full Model	221.84	143.19	234.02	181.56
	No residual topic	439.60	325.42	534.39	416.61
	ϵ constant	221.80	152.49	234.20	187.91
	No residual topic and ϵ constant	439.08	332.52	533.98	424.88
	Residual topics only	216.05	147.52	227.23	184.40
	LDA	441.42	334.93	533.59	427.27
Movies	Full Model	396.67	424.45	424.00	719.32
	No residual topic	652.54	624.30	824.28	897.53
	ϵ constant	398.68	480.64	424.00	851.97
	No residual topic and ϵ constant	652.62	802.93	822.93	991.29
	Residual topics only	386.44	463.65	408.43	892.43
	LDA	650.93	792.39	825.12	995.22
TV show episodes	Full Model	345.33	408.97	344.26	676.94
	No residual topic	513.58	664.27	512.64	767.39
	ϵ constant	345.84	481.06	344.42	935.64
	No residual topic and ϵ constant	512.75	1,100.60	511.32	1,170.60
	Residual topics only	339.95	457.47	338.45	928.63
	LDA	510.56	1,097.40	509.23	1,170.60

Note: fit and predictive performance are measured using perplexity (lower values indicate better fit).

tiveness measures between the version with 1,000 words and the one with 2,000 words is 0.79 for "inside the cone distinctiveness" ($p < 0.01$), 0.78 for "outside the cone distinctiveness" ($p < 0.01$), and 0.98 for "outside the cone emphasis in summary" ($p < 0.01$) in the marketing academic papers dataset. The correlations are respectively 0.56, 0.78-0.79 and 0.97 (all p's < 0.01) in the movies dataset (the set of movies included in the two regressions are not identical, hence correlations may vary slightly across the two regressions); and 0.81, 0.94 and 0.91 in the TV shows dataset (all p's < 0.01). We see that results are in general comparable to those obtained with 1,000 words in the vocabulary. "Inside the cone distinctiveness" and

"outside the cone distinctiveness" are directionally similar but become insignificant in the marketing academic papers dataset, and "inside the cone distinctiveness" becomes significant in the TV shows dataset.

Table WA15: Link Between Distinctiveness Measures and Citations - Marketing Academic Papers. 2,000 Words in Vocabulary.

Covariates	**DV=log(1+#citations)**
Journal fixed effects	✓
Publication year fixed effects	✓
Intensities of paper on non-flat regular topics	✓
Number of pages in paper	0.035**
"Inside the cone distinctiveness" (from journal)	0.042
"Outside the cone distinctiveness"	0.119
"Outside the cone emphasis in summary"	0.074**
Number of parameters	37
Number of observations	1,000
R^2	0.350

Note: OLS regression. *: significant at $p < 0.10$. **: significant at $p < 0.05$. All three distinctiveness measures are standardized across papers for interpretability.

Table WA17: Link Between Distinctiveness Measures and Performance - TV Episodes. 2,000 Words in Vocabulary.

Covariates	**DV=episode rating**
TV series fixed effects	✓
Intensities of script on non-flat regular topics	✓
"Inside the cone distinctiveness" (from TV series)	-0.042**
"Outside the cone distinctiveness"	0.083**
"Outside the cone emphasis in summary"	0.001
Number of parameters	348
Number of observations	9,285
R^2	0.688

Note: OLS regression. *: significant at $p < 0.10$. **: significant at $p < 0.05$. All three distinctiveness measures and episode ratings are standardized across episodes for interpretability.

Table WA16: Link Between Distinctiveness Measures and Performance - Movies. 2,000 Words in Vocabulary.

Covariates	**DV=movie rating**	**DV=log(return on investment)**
MPAA rating fixed effects	✓	✓
Genre fixed effects	✓	✓
Intensities of script on non-flat regular topics	✓	✓
Movie duration (in min)	0.002	-0.002
Log(inflation-adjusted production budget)	-0.088**	-0.307**
Movie rating	–	0.464**
"Inside the cone distinctiveness" (from genre)	-0.051	-0.118
"Outside the cone distinctiveness"	0.282**	-0.227
"Outside the cone emphasis in summary"	-0.001	0.001
Number of parameters	48	49
Number of observations	596	581
R^2	0.352	0.245

Note: each column corresponds to one regression estimated separately using OLS. *: significant at $p < 0.10$. **: significant at $p < 0.05$. All three distinctiveness measures and movie ratings are standardized across movies for interpretability. Observations in the first (resp., second) regression are limited to movies for which production budget was available (resp., production budget and box office performance were available).

H.1.2 500 Words in Vocabulary

Similarly, we decrease the size of the vocabulary to 500. Table WA18 reports the fit and predictive performance of the various benchmarks under this alternative vocabulary size. Results are qualitatively similar to those in Table 2.

Table WA18: Fit and Predictive Performance - Marketing Academic Papers. 500 Words in Vocabulary.

		Fit			Predictive perf.
	Approach	**Calibration documents**	**Calibration summaries**	**Validation documents**	**Validation summaries**
Marketing academic papers	Full Model	50.50	34.51	53.00	43.26
	No residual topic	82.98	59.41	90.46	67.66
	ϵ constant	50.52	37.67	53.01	43.65
	No residual topic and ϵ constant	83.17	62.62	90.75	72.20
	Residual topics only	49.42	36.71	51.59	43.05
	LDA	82.76	62.34	90.10	71.14
Movies	Full Model	75.33	77.49	80.46	134.20
	No residual topic	109.88	112.63	125.37	136.49
	ϵ constant	75.41	99.42	80.55	156.32
	No residual topic and ϵ constant	110.27	143.09	126.21	180.37
	Residual topics only	73.19	94.61	77.24	155.70
	LDA	108.24	145.62	125.15	184.01
TV show episodes	Full Model	163.46	137.15	163.52	228.03
	No residual topic	238.06	196.15	236.34	227.50
	ϵ constant	163.65	181.61	163.64	317.88
	No residual topic and ϵ constant	237.45	316.50	235.70	352.34
	Residual topics only	158.07	172.76	157.30	317.46
	LDA	234.48	309.55	232.37	349.94

Note: fit and predictive performance are measured using perplexity (lower values indicate better fit).

In the main model, we find 26 non-flat regular topics in the marketing academic papers datasets, 27 in the movies dataset, and 18 in the TV shows dataset. Tables WA19 to WA21 replicate Tables 4 to 6 with 500 of words in the vocabulary. The correlation in the distinctiveness measures between the version with 1,000 words and the one with 500 words is 0.89 for "inside the cone distinctiveness" ($p < 0.01$), 0.77 for "outside the cone

distinctiveness" ($p < 0.01$), and 0.98 for "outside the cone emphasis in summary" ($p < 0.01$) in the marketing academic papers dataset. The correlations are respectively 0.60-0.61, 0.85 and 0.96 (all p's < 0.01) in the movies dataset (the set of movies included in the two regressions are not identical, hence correlations may vary slightly across the two regressions); and 0.76, 0.87 and 0.91 in the TV shows dataset (all p's < 0.01). We see that results are in general comparable to those obtained with 1,000 words in the vocabulary. Results are directionally consistent with those with 1,000 words in the marketing academic papers dataset (although coefficients are not significant), results are similar in the movies dataset with "outside the cone emphasis in summary" becoming significant in the analysis of movie ratings and "inside the cone distinctiveness" becoming marginally significant in the analysis of ROI, and results are similar in the TV shows dataset.

Table WA19: Link Between Distinctiveness Measures and Citations - Marketing Academic Papers. 500 Words in Vocabulary.

Covariates	**DV=log(1+#citations)**
Journal fixed effects	✓
Publication year fixed effects	✓
Intensities of paper on non-flat regular topics	✓
Number of pages in paper	0.050**
"Inside the cone distinctiveness" (from journal)	0.036
"Outside the cone distinctiveness"	0.070
"Outside the cone emphasis in summary"	0.031
Number of parameters	39
Number of observations	1,000
R^2	0.336

Note: OLS regression. *: significant at $p < 0.10$. **: significant at $p < 0.05$. All three distinctiveness measures are standardized across papers for interpretability.

Table WA20: Link Between Distinctiveness Measures and Performance - Movies. 500 Words in Vocabulary.

Covariates	**DV=movie rating**	**DV=log(return on investment)**
MPAA rating fixed effects	✓	✓
Genre fixed effects	✓	✓
Intensities of script on non-flat regular topics	✓	✓
Movie duration (in min)	0.002	-0.001
Log(inflation-adjusted production budget)	-0.092**	-0.388**
Movie rating	–	0.433**
"Inside the cone distinctiveness" (from genre)	-0.025	-0.142*
"Outside the cone distinctiveness"	0.256**	-0.042
"Outside the cone emphasis in summary"	0.082**	0.055
Number of parameters	57	58
Number of observations	596	581
R^2	0.353	0.274

Note: each column corresponds to one regression estimated separately using OLS. *: significant at $p < 0.10$. **: significant at $p < 0.05$. All three distinctiveness measures and movie ratings are standardized across movies for interpretability. Observations in the first (resp., second) regression are limited to movies for which production budget was available (resp., production budget and box office performance were available).

Table WA21: Link Between Distinctiveness Measures and Performance - TV Episodes. 500 Words in Vocabulary.

Covariates	**DV=episode rating**
TV series fixed effects	✓
Intensities of script on non-flat regular topics	✓
"Inside the cone distinctiveness" (from TV series)	0.022
"Outside the cone distinctiveness"	0.082**
"Outside the cone emphasis in summary"	-0.003
Number of parameters	339
Number of observations	9,285
R^2	0.687

Note: OLS regression. *: significant at $p < 0.10$. **: significant at $p < 0.05$. All three distinctiveness measures and episode ratings are standardized across episodes for interpretability.

H.1.3 Simulation to Explore Sensitivity to Vocabulary Size

We see that as the vocabulary size changes, some coefficients in the regressions reported in this subsection sometimes lose or gain statistical significance (although we do not see any reversal, i.e., the same coefficient being statistically significant in opposite directions across vocabulary sizes). We run a simulation study to explore how the correlation between the "true" measures of distinctiveness and their estimates is affected by vocabulary size. In particular, we simulate situations in which the vocabulary is larger than needed and "irrelevant" words are included in the vocabulary, as well as situations in which the vocabulary is smaller than it should and "relevant" words are omitted. We use the marketing academic papers as our dataset of reference. We assume that the data generating process follows the assumptions of the model and use the original model estimates from this dataset as the true parameters, but we add irrelevant words and vary the number of words in the vocabulary used for model estimation.

More precisely, we assume 1,000 words in the "true" vocabulary and 1,000 documents in the dataset, and assume that the true parameters $\{\beta_{kv}^{reg}, \epsilon_k^{reg}, \beta_{dv}^{res}, \epsilon_d^{res}, \theta_{dk}^{reg}\}_{d,k,v}$ are equal to the model estimates from the marketing academic papers dataset reported in the main paper. We generate simulated word counts for all documents and summaries according to the model's data generating process: $w_{dv} \sim Poisson(\sum_k \theta_{dk}^{reg}\beta_{kv}^{reg} + \beta_{dv}^{res})$; $w_{dv}^{summary} \sim Poisson(\sum_k \theta_{dk}^{reg}\beta_{kv}^{reg}\epsilon_k^{reg} + \beta_{dv}^{res}\epsilon_d^{res})$. Note that drawing word occurrences from a Poisson distribution introduces a first source of noise into the data.

Next, we introduce another set of 1,000 irrelevant words that appear randomly across documents, with no structure. We simulate the number of occurrences of each irrelevant word v' in each document d and the summary of that document by drawing from Poisson distributions: $w_{dv'} \sim Poisson(\lambda')$; $w_{dv'}^{summary} \sim Poisson(\lambda'')$. We set λ' (respectively, λ'') so that the average number of occurrences of irrelevant words across documents (respectively, summaries) is 19% of the average number of occurrences of relevant words. That is, the average of w_{dv} (respectively, $w_{dv}^{summary}$) across words and documents is equal to the average of $w_{dv'}$ (respectively, $w_{dv'}^{summary}$), multiplied by 0.19. We choose 0.19 because this is the proportion we observe in the marketing academic papers when comparing the average number of occurrences of the top 1,000 words to the next 1,000 words, when words are ranked according to $tf\text{-}idf$.

We calibrate three versions of the model. In the first version, we include all 2,000 words into the vocabulary (i.e., all relevant and all irrelevant words are included in the vocabulary when estimating the model). In the second version, we select 1,000 words using $tf\text{-}idf$ (i.e., the number of words in the vocabulary is the same as the true number of relevant words, although the exact sets of words may be different). In the third version, we select 500 words

using tf-idf (i.e., the number of words in the vocabulary is lower than the true number of relevant words).

We calibrate each version of the model and compute the three distinctiveness measures obtained using the new model estimates. For each of the three measures, we compute the correlation across documents between the true and estimated distinctiveness (where the true measures are the ones based on the "true" model parameters). We repeat the exercise 100 times, i.e., with 100 different simulated draws of word occurrences.

Table WA22 reports the average and standard deviation (across the 100 replications) of the correlations for each version of the model and each distinctiveness measure. First, we see that that when the number of words in the vocabulary coincides with the true number of relevant words, the measures are well recovered, with the correlations between the true and estimated measures ranging from 0.868 to 0.917. Next, we see that these correlations tend to drop when irrelevant words are added into the vocabulary (2,000 words in total), or when the vocabulary size is too small (500 words). That is, the distinctiveness measures become less precise, which could explain why some distinctiveness measures that are significant when 1,000 words are included in the vocabulary, become insignificant under different vocabulary sizes.

Table WA22: Correlation Between True and Estimated Distinctiveness. Simulation Results with Varying Vocabulary Sizes.

	2,000 words	**1,000 words**	**500 words**
"Inside the cone distinctiveness" (from journal)	0.770 (0.070)	0.868 (0.029)	0.879 (0.016)
"Outside the cone distinctiveness"	0.860 (0.032)	0.894 (0.021)	0.833 (0.028)
"Outside the cone emphasis in summary"	0.819 (0.014)	0.917 (0.009)	0.896 (0.013)

Note: average from 100 replications, with standard deviation in parentheses.

H.2 Alternative Measures of Performance

H.2.1 DIC

As an alternative measure of model performance, we compute the Deviance Information Criterion or DIC (Celeux et al., 2006; Spiegelhalter et al., 2002). To compute the DIC, we draw 1,000 sets of parameters from the posterior distribution conditional on the data (given by our Variational Inference algorithm), and estimate the DIC using Equation (2) in Celeux et al. (2006), taking the expectation over posterior draws. Table WA23 reports the results. It is important to keep in mind that the DIC measure for LDA should *not* be directly compared to the DIC measures for Poisson factorization. That is because the two models assume different likelihood functions. While LDA takes the number of tokens as given and predicts the word corresponding to each token, Poisson factorization also predicts the number of tokens. As explained in the paper, perplexity allows us to compare the two approaches directly, by transforming the Poisson rates into multinomial distributions.

Table WA23: DIC.

		DIC ($\times 10^6$)
Marketing academic papers	Full Model	1.919
	No residual topic	12.656
	ϵ constant	1.927
	No residual topic and ϵ constant	7.188
	Residual topics only	1.934
	LDA	2.310
Movies	Full Model	1.426
	No residual topic	5.762
	ϵ constant	1.494
	No residual topic and ϵ constant	4.050
	Residual topics only	1.477
	LDA	11.645
TV show episodes	Full Model	32.639
	No residual topic	185.842
	ϵ constant	32.876
	No residual topic and ϵ constant	94.956
	Residual topics only	33.838
	LDA	189.709

Note: Lower values are preferred.

H.2.2 Alternative Measures of Predictive Performance

We consider an alternative test of predictive performance. The test reported in the paper focuses on predicting the content of the validation summaries on the basis of the validation documents. As an alternative, we consider the task of predicting some of the word occurrences in each validation document, on the basis of a subset of this document. For each validation document, we randomly holdout 20% of the observed word occurrences. Intuitively, this task would correspond to randomly "erasing" 20% of the document, and predicting the part that was erased based on the part that was not erased. We compute the

perplexity score on the heldout word occurrences as our measure of predictive performance on validation documents. In addition, we predict the content of each validation summary on the basis of the subset of the full document that was not "erased," and report the perplexity score as our measure of predictive performance on validation summaries. Table WA24 reports the results. We see that the full model performs best in terms of predicting the heldout portion of each validation document and predicting its summary, based on a subset of the full document. The exception is the marketing academic papers dataset in which the "Residual topics only" version performs slightly better at predicting the heldout portion of validation documents.

Table WA24: Fit and Predictive Performance. Alternative Measure of Predictive Performance.

		Fit			Predictive perf.	
		Calibration documents	**Calibration summaries**	**Validation documents**	**Validation documents**	**Validation summaries**
Marketing academic papers	Full Model	104.54	67.60	110.47	129.39	85.34
	No residual topic	197.04	141.39	227.56	232.60	170.51
	ϵ constant	104.46	73.13	110.49	129.40	88.56
	No residual topic and ϵ constant	196.91	146.39	227.44	232.12	177.54
	Residual topics only	101.83	70.84	107.06	128.42	87.11
	LDA	197.13	145.73	226.24	231.76	176.88
Movies	Full Model	168.72	177.28	178.74	233.12	301.08
	No residual topic	265.80	279.76	323.26	336.48	355.68
	ϵ constant	169.07	213.13	178.78	233.54	356.57
	No residual topic and ϵ constant	265.50	346.99	323.23	337.01	430.03
	Residual topics only	163.82	204.87	171.59	233.28	363.48
	LDA	267.29	343.28	326.81	340.02	425.94
TV show episodes	Full Model	241.61	246.36	237.25	370.72	420.34
	No residual topic	361.10	416.28	357.48	381.23	440.59
	ϵ constant	241.58	311.23	237.09	370.75	590.37
	No residual topic and ϵ constant	360.29	633.89	356.94	380.88	690.57
	Residual topics only	234.86	294.33	230.10	387.26	588.57
	LDA	360.56	643.63	356.64	380.95	701.21

Note: fit and predictive performance are measured using perplexity (lower values indicate better fit).

H.3 Link Between Distinctiveness and Success: Alternative Specifications

We test alternative specifications in the analysis of movies. In one specification, we remove movie rating as a predictor of log(return on investment). In another, we replace the DV with the log of the inflation-adjusted box office performance. We run this latter specification with and without the movie rating as a predictor. Consistent with Table 5, we find no significant relation between any of the distinctiveness measures and financial performance.

Table WA25: Link Between Distinctiveness Measures and Performance - Movies. Alternative Specifications.

Covariates	**DV=log(return on investment)**	**DV=log(box office perf.)**	**DV=log(box office perf.)**
MPAA rating fixed effects	✓	✓	✓
Genre fixed effects	✓	✓	✓
Intensities of script on non-flat regular topics	✓	✓	✓
Movie duration (in min)	-0.001	-0.003	-0.001
Log(inflation-adjusted production budget)	-0.391**	0.671**	0.609**
Movie rating	–	0.451**	–
"Inside the cone distinctiveness" (from genre)	-0.070	-0.027	-0.070
"Outside the cone distinctiveness"	0.003	-0.109	0.003
"Outside the cone emphasis in summary"	0.093	0.069	0.093
Number of parameters	54	55	54
Number of observations	581	581	581
R^2	0.198	0.448	0.401

Note: each column corresponds to one regression estimated separately using OLS. *: significant at $p < 0.10$. **: significant at $p < 0.05$. All three distinctiveness measures and movie ratings are standardized across movies for interpretability. Observations are limited to movies for which production budget and box office performance were available.

H.4 Alternative Measure of "Inside the Cone Distinctiveness"

In the analyses reported in Tables 4 to 6, "inside the cone distinctiveness" is based on comparing each document to a relevant subgroup (e.g., genre). Here, we replicate the analysis with a measure of "inside the cone" distinctiveness that compares each document to the entire set of training documents. That is, we measure the distinctiveness of a document d with intensities on regular topics $\{\theta_{dk}^{reg}\}_k$, as: $1-\Sigma_k \frac{|\theta_{dk}^{reg}-\theta_k^{reg}|}{\theta_{dk}^{reg}+\theta_k^{reg}+0.0001}$, where θ_k^{reg} is the average intensity on topic k across all documents in the training sample. Tables WA26 to WA28 report the results. Results are similar to the ones obtained with the measure that compares each document to its reference group.

Table WA26: Link Between Distinctiveness Measures and Citations - Marketing Academic Papers. "Inside the Cone Distinctiveness" Based on Entire Set of Training Documents.

Covariates	**DV=log(1+#citations)**
Journal fixed effects	✓
Publication year fixed effects	✓
Intensities of paper on non-flat regular topics	✓
Number of pages in paper	0.039**
"Inside the cone distinctiveness"	0.086**
"Outside the cone distinctiveness"	0.148**
"Outside the cone emphasis in summary"	0.062**
Number of parameters	43
Number of observations	1,000
R^2	0.355

Note: OLS regression. *: significant at $p < 0.10$. **: significant at $p < 0.05$. All three distinctiveness measures are standardized across papers for interpretability.

Table WA27: Link Between Distinctiveness Measures and Performance - Movies. "Inside the Cone Distinctiveness" Based on Entire Set of Training Documents.

Covariates	**DV=movie rating**	**DV=log(return on investment)**
MPAA rating fixed effects	✓	✓
Genre fixed effects	✓	✓
Intensities of script on non-flat regular topics	✓	✓
Movie duration (in min)	0.003*	-0.003
Log(inflation-adjusted production budget)	-0.094**	-0.330**
Movie rating	–	0.453**
"Inside the cone distinctiveness"	-0.097**	-0.011
"Outside the cone distinctiveness"	0.252**	-0.101
"Outside the cone emphasis in summary"	0.052	0.071
Number of parameters	54	55
Number of observations	596	581
R^2	0.358	0.261

Note: each column corresponds to one regression estimated separately using OLS. *: significant at $p < 0.10$. **: significant at $p < 0.05$. All three distinctiveness measures and movie ratings are standardized across movies for interpretability. Observations in the first (resp., second) regression are limited to movies for which production budget was available (resp., production budget and box office performance were available).

Table WA28: Link Between Distinctiveness Measures and Performance - TV Episodes. "Inside the Cone Distinctiveness" Based on Entire Set of Training Documents.

Covariates	**DV=episode rating**
TV series fixed effects	✓
Intensities of script on non-flat regular topics	✓
"Inside the cone distinctiveness"	0.001
"Outside the cone distinctiveness"	0.076**
"Outside the cone emphasis in summary"	-0.007
Number of parameters	340
Number of observations	9,285
R^2	0.687

Note: OLS regression. *: significant at $p < 0.10$. **: significant at $p < 0.05$. All three distinctiveness measures and episode ratings are standardized across episodes for interpretability.

H.5 Capturing Uncertainty in Distinctiveness Measures

Tables 4 to 6 are based on distinctiveness measures that are computed using point estimates of the model parameters. In order to reflect uncertainty in these measures, we run each regression 1,000 times, each time based on a different random draw from the posterior distribution of the model parameters. That is, in each iteration we draw from the posterior distribution of the model parameters, compute the distinctiveness measures, and run the regression. Tables WA29 to WA31 report the average coefficients across the 1,000 iterations. For each coefficient, we also report whether the 90% and 95% credible intervals across iterations include 0. We find that results are overall consistent with those reported in Tables 4 to 6, with some additional coefficients becoming "significant" (i.e., the 95% credible interval does not include 0).

Table WA29: Link Between Distinctiveness Measures and Citations - Marketing Academic Papers. Multiple Draws from Posterior Distribution.

Covariates	**DV=log(1+#citations)**
Journal fixed effects	✓
Publication year fixed effects	✓
Intensities of paper on non-flat regular topics	✓
Number of pages in paper	0.041**
"Inside the cone distinctiveness" (from journal)	0.057**
"Outside the cone distinctiveness"	0.126**
"Outside the cone emphasis in summary"	0.054**
Number of parameters	43
Number of observations	1,000
R^2	0.352

Note: average results across 1,000 OLS regressions, each based on a random draw from the posterior distribution of the model parameters. *: 90% credible interval across iterations does not include 0. **: 95% credible interval across iterations does not include 0.

Table WA30: Link Between Distinctiveness Measures and Performance - Movies. Multiple Draws from Posterior Distribution.

Covariates	**DV=movie rating**	**DV=log(return on investment)**
MPAA rating fixed effects	✓	✓
Genre fixed effects	✓	✓
Intensities of script on non-flat regular topics	✓	✓
Movie duration (in min)	0.003**	-0.003**
Log(inflation-adjusted production budget)	-0.093**	-0.329**
Movie rating	–	0.451**
"Inside the cone distinctiveness" (from genre)	-0.086**	-0.027
"Outside the cone distinctiveness"	0.239**	-0.109**
"Outside the cone emphasis in summary"	0.051**	0.069**
Number of parameters	54	55
Number of observations	596	581
R^2	0.356	0.262

Note: average results across 1,000 OLS regressions, each based on a random draw from the posterior distribution of the model parameters. *: 90% credible interval across iterations does not include 0. **: 95% credible interval across iterations does not include 0.

Table WA31: Link Between Distinctiveness Measures and Performance - TV Episodes. Multiple Draws from Posterior Distribution.

Covariates	**DV=episode rating**
TV series fixed effects	✓
Intensities of script on non-flat regular topics	✓
"Inside the cone distinctiveness" (from TV series)	-0.002
"Outside the cone distinctiveness"	0.058**
"Outside the cone emphasis in summary"	-0.007**
Number of parameters	340
Number of observations	9,285
R^2	0.687

Note: average results across 1,000 OLS regressions, each based on a random draw from the posterior distribution of the model parameters. *: 90% credible interval across iterations does not include 0. **: 95% credible interval across iterations does not include 0.

H.6 Distinctiveness Measures Based on Standard Topic Models

We test the link between success and measures of distinctiveness based on standard models. Standard models like LDA or Poisson Factorization do not include residual topics and do not model the link between full documents and their summaries. Accordingly, only one of our three distinctiveness measures, "inside the cone distinctiveness," is available for these models. We regress success on this distinctiveness measure, intensities on standard topics, and the same additional covariates as in Tables 4 to 6. Results are reported in Tables WA32 to WA34. We see that the number of parameters and the R^2 are higher in these tables compared to Tables 4 to 6, as the number of non-flat topics is larger for the standard models that do not contain residual topics. "Inside the cone distinctiveness" is never significantly related to success when computed using these standard models, with the exception of Table WA33b in which the link with log(return on investment) is marginally significant.

Table WA32: Link Between Distinctiveness and Citations - Marketing Academic Papers. Distinctiveness Measures Based on Standard Topic Models.

(a) LDA

Covariates	**DV=log(1+#citations)**
Journal fixed effects	✓
Publication year fixed effects	✓
Intensities of paper on non-flat regular topics	✓
Number of pages in paper	0.048**
"Inside the cone distinctiveness" (from journal)	-0.006
Number of parameters	110
Number of observations	1,000
R^2	0.472

(b) Standard Poisson Factorization

Covariates	**DV=log(1+#citations)**
Journal fixed effects	✓
Publication year fixed effects	✓
Intensities of paper on non-flat regular topics	✓
Number of pages in paper	0.049**
"Inside the cone distinctiveness" (from journal)	0.031
Number of parameters	111
Number of observations	1,000
R^2	0.446

Note: OLS regression. *: significant at $p < 0.10$. **: significant at $p < 0.05$. "Inside the cone distinctiveness" is standardized across papers for interpretability.

Table WA33: Link Between Distinctiveness Measures and Performance - Movies. Distinctiveness Measures Based on Standard Topic Models.

(a) LDA

Covariates	**DV=movie rating**	**DV=log(return on investment)**
MPAA rating fixed effects	✓	✓
Genre fixed effects	✓	✓
Intensities of script on non-flat regular topics	✓	✓
Movie duration (in min)	0.002	-0.001
Log(inflation-adjusted production budget)	-0.032	-0.303**
Movie rating	–	0.438**
"Inside the cone distinctiveness" (from genre)	-0.070	-0.044
Number of parameters	127	128
Number of observations	596	581
R^2	0.461	0.383

(b) Standard Poisson Factorization

Covariates	**DV=movie rating**	**DV=log(return on investment)**
MPAA rating fixed effects	✓	✓
Genre fixed effects	✓	✓
Intensities of script on non-flat regular topics	✓	✓
Movie duration (in min)	0.002	-0.002
Log(inflation-adjusted production budget)	-0.055	-0.341**
Movie rating	–	0.470**
"Inside the cone distinctiveness" (from genre)	-0.013	-0.154*
Number of parameters	128	129
Number of observations	596	581
R^2	0.416	0.406

Note: each column corresponds to one regression estimated separately using OLS. *: significant at $p < 0.10$. **: significant at $p < 0.05$. "Inside the cone distinctiveness" and movie ratings are standardized across movies for interpretability. Observations in the first (resp., second) regression are limited to movies for which production budget was available (resp., production budget and box office performance were available).

Table WA34: Link Between Distinctiveness Measures and Performance - TV Episodes. Distinctiveness Measures Based on Standard Topic Models.

(a) LDA

Covariates	**DV=episode rating**
TV series fixed effects	✓
Intensities of script on non-flat regular topics	✓
"Inside the cone distinctiveness" (from TV series)	0.027
Number of parameters	418
Number of observations	9,285
R^2	0.692

(b) Standard Poisson Factorization

Covariates	**DV=episode rating**
TV series fixed effects	✓
Intensities of script on non-flat regular topics	✓
"Inside the cone distinctiveness" (from TV series)	-0.023
Number of parameters	419
Number of observations	9,285
R^2	0.691

Note: OLS regression. *: significant at $p < 0.10$. **: significant at $p < 0.05$. "Inside the cone distinctiveness" and episode ratings are standardized across episodes for interpretability.

References

David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent dirichlet allocation. *Machine Learning*, 3(4/5):993–1022, 2003.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *arXiv preprint arXiv:1601.00670*, 2016.

Gilles Celeux, Florence Forbes, Christian P. Robert, and D. Michael Titterington. Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4):651–673, 2006.

Prem Gopalan, Jake M Hofman, and David M Blei. Scalable recommendation with poisson factorization. *arXiv preprint arXiv:1311.1704*, 2013.

Prem K Gopalan, Laurent Charlin, and David Blei. Content-based recommendations with poisson factorization. In *Advances in Neural Information Processing Systems*, pages 3176–3184, 2014.

David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639, 2002.