Modeling Virtual Organizations with Latent Dirichlet Allocation: A Case for Natural Language Processing

Alexander J. Gross (U. of Maine) and Dhiraj Murthy (Goldsmiths College

Accepted and will appear in an issue of *Neural Networks at*
http://www.sciencedirect.com/science/article/pii/S0893608014001075

Abstract

This paper explores a variety of methods for applying the Latent Dirichlet Allocation (LDA) automated topic modeling algorithm to the modeling of the structure and behavior of virtual organizations found within modern social media and social networking environments. As the field of Big Data reveals, an increase in the scale of social data available presents new challenges which are not tackled by merely scaling up hardware and software. Rather, they necessitate new methods and, indeed, new areas of expertise. Natural language processing provides one such method. This paper applies LDA to the study of scientific virtual organizations whose members employ social technologies. Because of the vast data footprint in these virtual platforms, we found that natural language processing was needed to 'unlock' and render visible latent, previously unseen conversational connections across large textual corpora (spanning profiles, discussion threads, forums, and other social media incarnations). We introduce variants of LDA and ultimately make the argument that natural language processing is a critical interdisciplinary methodology to make better sense of social 'Big Data' and we were able to successfully model nested discussion topics from forums and blog posts using LDA. Importantly, we found that LDA can move us beyond the state-of-the-art in conventional Social Network Analysis techniques.

Keywords: natural language processing, Latent Dirichlet Allocation, Big Data, social media, virtual organizations

1. Introduction

In recent years, the Internet has undergone enormous transformations. From its inception as a framework for the interconnection of fragments of information from disparate locations through a vast network of hyperlinks, the Internet has evolved into a new medium of communication that almost seamlessly connects individuals with one another. Far from early critiques that the Internet separates and isolates people, it could now be argued that an important function the Internet is in fact to keep people socially connected and even increase their social capital (Wellman, Haase et al. 2001). With sites including Facebook and Twitter amongst the most heavily used sites on the Internet, and even sites traditionally thought of as informational like Google relying heavily on social features, social technologies have become increasingly important to us. Indeed, the Internet has become primarily semantic, contextual, and social (Gruber 2008). A key challenge now is to parse, understand, and visualize these online formations and spaces and their role in our lives.

Many social networking sites implement features which allow users to record and report their relationships with other users and groups (e.g. friending, liking, and following) (boyd and Ellison 2007). A whole research domain has arisen in recent years to quantitatively explore the kinds of networks defined by such user-reported relationships. But a complete and thorough analysis of online communities and organizations, which stops there, would be incomplete. These types of relationships only capture a small portion of the activity that defines any true online community. For example, virtual organizations, "'collection[s] of geographically dispersed individuals, groups, organizational units - either belonging or not belonging to the same organization - or entire organizations that depend on electronic links in order to complete the production process" (Travica 1997), involve complex social webs which depend on the development and maintenance of trust. And because virtual organizations are increasing mediated by social technologies, they often include virtual spaces or forums for fostering trust and completing tasks. However, an analysis of these communicative spaces alone is insufficient to the task of understanding the organizational and community-building potential provided by modern virtual social formations. Understanding the human relationships within virtual organizations requires much more than capturing who interacts with whom and the topics about which people communicate. Rather, the formations of profiles and other facets of people's virtual presence are negotiated and constructed over time in complex ways. Understanding this activity within virtual organizations involves the study of many users, actions, connections, and communities taking place across a universe of diverse threads, discussions, groups, and micro-communities within a typical social platform and potentially even across multiple social technologies.

In the last decade, even as interactive social technologies were settling in as the dominant online paradigm, there was a severe lack of accepted research methods or even availability of information about fledgling online social activity. Even where rich information was available, robust computational analysis of that information was often intractable. This status quo often led to favoring approaches to network analysis focused on the analysis of reported relationships, as there was an established framework for the study of such structures borrowed from graph theory (Fombrun 1982). These methods are and remain

important to understanding social networks online. However, they can and should be augmented with other types of data and methods of analysis. Natural language processing provides one important avenue for this.

In an earlier issue of this Journal, Fabish et al. (2013) argue that "Artificial intelligence is facing real world problems and thus machine learning problems become more and more complex." Part of this real world problem is the exponentially increasing volume of data that machine learning has to process. The debates around 'Big Data' challenges have appeared in this Journal as well as across the computer science literature more broadly (Kantardzic 2011). A particular challenge for artificial intelligence in the context of Big Data is to ensure less—not more—effort is "shifted from human to machine" (Fabisch, Kassahun et al. 2013).

Of course, not all Big Data is equal. A major challenge facing the artificial intelligence community is machine learning with small chunks of text, such as text commonly found on online social networks, social media, and other online spaces. For example, Twitter data has a wide range of text quality, text length, and content types. Machine learning with 140-character tweets is possible and has been done by many (Pak and Paroubek 2010), but the task remains rife with problems, especially when conventional machine learning algorithms are applied. As Cambria and White (2014) argue, "NLP research has evolved from the era of punch cards and batch processing […] to the era of Google," and machine learning continues to evolve to adapt to complex settings, such as the analysis of social media data. One part of this evolution is the move from coarse to fine-grained analysis methods. As Cambria et al. (2013) point out in their discussion of opinion mining and sentiment analysis, early NLP methods often classified sentiment based on a whole document, whereas newer methods strive to analyze segment-level sentiment.

In this article, we describe application of one of these more fine-grained NLP techniques—Latent Dirichlet Allocation (LDA) (Blei, Ng et al. 2003) —to information gathered from two prominent virtual communities of life scientists. LDA is a robust and versatile unsupervised topic modeling technique, originally developed to identify latent topics within a collection of text documents. It has shown great

flexibility in being easily adapted to situations where objects in a collection are each associated with a unique set of exchangeable attributes (words, in the case of text documents). In general, the technique discovers latent topics of associated and co-occurring attributes within the collection. A latent topic has a probability distribution over words (as opposed to a strict list of words that are included in or excluded from the topic). Instead of simply determining an object's simple group membership, as is the case with many machine learning algorithms, LDA uses a mixture model that models each of its objects as drawing proportionally from each of a set of latent topics. These models provide a rich, almost genome-like structure for the comparison of objects, classifying each on the entire range of latent groupings. This article first introduces the context of our research question, then describes LDA and its variants before moving to our study methods and results.

2. Background

In this section, we will provide a brief background to the LDA method for topic modeling and highlight some of the opportunities that exist in the application of LDA to understanding collective behavior and latent, normally unseen network structures that can be discovered and explored from aggregated communication (in our case, drawn from virtual communities and organizations). Of particular interest to our research are the patterns and networks of latent behavior and communication that help to understand and illuminate the collective activity of scientific social networking sites and particularly the virtual organizations that develop from them.

Social media and social networking technologies have become ubiquitous in our social lives. However, they are also increasingly pervasive in organizational settings. For example, corporate internal social media systems such as HP's WaterCooler (Brzozowski 2009) and IBM's Beehive (Geyer, Dugan et al. 2008) confirm the utility of social technologies to organizational innovation, collaboration and general knowledge sharing. Individual discussion threads and even small clusters of interactions on these platforms can be readily analyzed, but it is not easy or straightforward to do this on a much larger scale. As the field of Big Data reveals, an increase in the scale of social data available cannot be effectively managed by merely scaling up hardware and software, but creates new challenges which necessitate new

methods and, indeed, new areas of expertise (Kaisler, Armour et al. 2013). Our project is particularly interested in the study of virtual organizations mediated by social technologies.

2.1 Virtual Organizations

Virtual organizations (VOs) are organizations or enterprises not tied to a singular physical locality (i.e. a specific lab or work place), and are a product of changes in global economic, social, and political systems. A useful working definition of VOs is provided by Travica (1997) who views them as manifesting themselves as a "collection of geographically dispersed individuals, groups, organizational units - either belonging or not belonging to the same organization - or entire organizations that depend on electronic links in order to complete the production process." The work of Travica (1997) and Mowshowitz (1997), though useful in defining elements of VOs and mapping their history, does not offer a general articulation of what constitutes a virtual organization. Indeed, VOs are conceptualized differently in different contexts. The VOs form, disband, and re-configure as required for the task. A VO in this context is a virtual collection of geographically disparate team members brought together to solve a particular problem/task or accomplish a specific goal. Ultimately, in global virtual teams, the 'grid' is distributed human resources connected together through collaborative new media technologies to work together as a VO. In this way, VOs share with offline organizations a purpose of organizing individuals towards a common cause. But, with the exponential increases in textual data being produced with larger VO size, time and the technological capabilities of modern society, it can often be difficult to see the ways in which that common cause is being achieved.

Virtual organizations can be large and their data footprint is much larger. Indeed, in the two scientific platforms we studied, we not only observed virtual organizations of varying sizes, but witnessed a heterogeneity of interactions between the two. We used social network analysis (SNA) (Scott and Carrington 2011) to discern specific clusters, cliques, and other groupings within the two scientific platforms. A network-based approach was very useful in rendering visible networks of users and the ways in which they were connected. However, we discovered this was a very partial portrait. Specifically, it

prioritized particular types of users such as "leaders"/"brokers" within the network. It also did not speak to the role of topics within these network structures. Natural language processing presents an excellent means to resolve this dilemma.

3. Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a type of generative probabilistic model. As Cambria and White (2014) highlight, LDA is an endogenous NLP technique, which "involves the use of machine-learning techniques to perform semantic analysis of a corpus by building structures that approximate concepts from a large set of documents" without relying on external knowledge bases. LDA is a latent variable model in which each item in a collection (e.g., each text document in a corpus) is modeled as a finite mixture over an underlying set of topics. Each of these topics is characterized by a distribution over item properties (e.g., words). LDA assumes that these properties are exchangeable (i.e., ordering of words is ignored, as in many other "bag of words" approaches to text modeling), and that the properties of each document are observable (e.g., the words in each document are known). The word distribution for each topic and the topic distribution for each document are unobserved; they are learned from the data. The details of the model formulation and training procedure are described in the following subsection.

Once a model has been learned from a corpus, the topic distribution associated with each document can be represented as a vector, which can be used to calculate a distance between documents that is informative of their similarity. That is, documents that are similar will have similar distributions over topics, and thus be closer together in this vector space. This vector representation and notion of distance can therefore provide a foundation to classify, group, and identify relationships between documents.

LDA has proven useful for encapsulating and generating knowledge from large corpora, which in many cases were resistant or intractable to previous attempts to model the latent structure or relationships within the data (Blei, Ng et al. 2003). LDA has been applied to many types of problems, including modeling scientific digital library collections (Mann, Mimno et al. 2006), relationships between images and their captions (Blei & Jordan 2002) and topics within disasters from Twitter data (Kireyev, Palen et

al. 2009). It has been shown to perform as well or better than many other popular techniques for machine learning, data mining, and supervised and unsupervised classification of data. Indeed, LDA has been found to have a similar running time for processing as *k*-means (Wei and Croft 2006), a long-used approach for unsupervised clustering, which lacks LDA's capability to associate documents with a distribution over topics rather than assignment of each document to a single, unique topic. Modifications, extensions, improvements, and additions to LDA are being developed and released at a rapid pace; some relevant extensions are discussed later in this article.

3.1 LDA Topic Modeling

LDA models the relationships between words, documents, and topics in a corpus via a generative probabilistic model. Within this model documents are modeled as mixtures over latent topics, and each topic is modeled as an unique distribution over the entire observed vocabulary of the corpus. LDA makes the assumption that documents were generated via the following generative process described by Blei (2012):

1. For each Document randomly choose a distribution over topics.
2. For each word in the document
(a) Randomly choose a topic from the distribution over topics in step #1.
(b) Randomly choose a word from the corresponding distribution over the vocabulary.

Blei et al. (2003) provide the following detailed process:

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:
(a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
(b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Step 1 is not highly relevant to the process of determining the actual topic structures. Specifically, this

step involves imagining some random document of a length of words drawn from a Poisson distribution, but it does not figure in the actual reversal of the process. These process descriptions are describing an assumption for how a corpus of documents is formed (namely the 'bag of words' and topic mixture assumptions). The documents are in fact written by various authors, but it is the assumption of this generative process for some random θ (per-document topic distribution) and some random β (per topic word distribution) for the creation of documents that allows for the derivation of an equation to reverse the generative process given some provided evidence (the actual corpus of documents).

Blei et al. (2003) detail the derivation given the above assumptions. Equation 1 is the equation for the Dirichlet distribution.

$$p(\theta \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1},$$

Equation 1 from Blei et al. (2003)

This leads to Equation 2 which is (for a specific document) the joint distribution of θ (the topic mixture), z (some set of topic distributions), and w (the N words in that document) given α and β.

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta),$$

Equation 2 from Blei et al. (2003)

Equation 3 shows the marginal distribution of a document or the probability of some set of words w appearing in a document theoretically created via the above process and defined by α and β.

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta.$$

Equation 3 from Blei et al. (2003)

Equation 4 takes the product of the marginal distributions of each document to obtain an equation for the conditional probability of a corpus given α and β.

$$p(D|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d.$$

Equation 4 from Blei et al. (2003)

Theoretically, one could use Equation 4 and try all possible α and β to maximize the probability of a specific corpus (the one selected to be analyzed) given the specific evidence (words to document assignment existed in the provided corpus). It is this process which is intractable to compute.

3.2 α and β hyperparameters

α parameter is simply a number > 0 which is used in Dirichlet distribution. Dirichlet distributions give a k-dimensional distribution. Basically, a Dirichlet distribution allows one to select a random distribution, but one which is characterized by α. An α close to 1 is almost uniform, though each selection from the Dirichlet will be a unique random distribution. At α close to 0, the distributions are very unbalanced, meaning most of the weight of the distribution will be assigned to very few of the topics (similar to an exponential distribution). In other words, α determines the level of topic mixtures. At α close to 1, the mixture would be fairly uniform. At an α close to 0, the mixtures would be very "topic-y", meaning almost all of the probability would be concentrated in 1 or 2 of K topics.

β is a matrix of the probability of words to topics so the size is k-topics x length (vocabulary). The vocabulary is quite large for any corpus this is the part that is intractable to compute. There is no way to test every possible vocabulary to topic distribution to see which has the greatest probability given the provided corpus. Since p(D|α,β) of a corpus given α and β cannot be computed for every possible α and β to see which has the greatest probability given D (corpus of documents), it has to be estimated using Gibbs sampling (Geman and Geman 1984) or variation inference.

3.3 Variational Inference

The variational inference is the estimation of α and β given the evidence (provided corpus). The process is a two-step iterative process. In the first step (E(stimation)-step) for each document and some initial α and β, an optimization algorithm is solved to obtain a lower bound on the log-likelihood given the current α and β. The optimization problem yields two variational parameters, γ and φ, that give the tightest possible lower bound on the log-likelihood. The second step (M(aximization)-Step) attempts to maximize this lower bound with respect to α and β. Within this step, γ and φ are updated to maximize the lower bound on the log-likelihood. These two steps are repeated until the Kullback-Leibler divergence between the variational distribution and the "true" posterior distribution converge below some threshold. At this point, we can say we have approximated α and β (the "true posterior distribution").

This process can be understood as a graphical model represented with "plate" notation (Buntine 1994), as shown in Figure 1. In this notation, a repeating group of model variables is drawn within a rectangular plate, and the number of repetitions is indicated on the plate. In Figure 1, the N plate is the collection of words in a given document and the collection of all topic vocabulary distribution and the D plate represents the collection of documents within the collection.

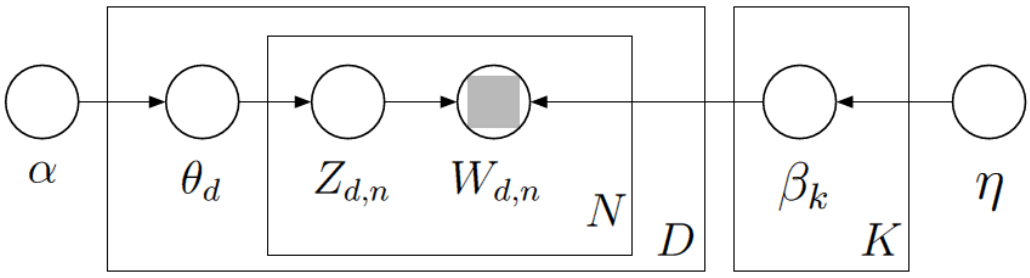

Figure 1: Graphical Model Representation of LDA (Blei 2012)

In Figure 1, the words w_{dn} are the only observable variables. Z_{dn} is the topic assignment of a given word in a given document. The values of α and β are hyperparameters that are set manually, as is the value of *K*, the number of topics. The theoretical distribution space of possible vocabulary to topic assignments is represented by Ƞ. Other variables - each topic's distribution over words, each document's distribution over topics, and the topic membership of each word - must be inferred from the data. The process for inferring these values is derived from methods developed from generative probabilistic modeling. In such a system, the observed evidence (or words in documents) and the hidden variable or latent topic structure are used to develop a joint probability distribution over all model variables. This joint distribution is used to compute the conditional or posterior distribution for the variables given the observed documents. The goodness of fit of the model can be tuned in several ways, including adjustment of the hyperparameters or number of topics, as well as limiting the number of iterations to perform in the estimation.

3.4 Utility

The LDA model has a number of notable benefits. First, it technically need not be applied to language at all, but to any entities in a collection that are composed of a set of exchangeable observations. For example, LDA has been applied to estimation of genetic ancestries through collections of expressed genes in individuals (Pritchard, Stephens et al. 2000), as well as to digital image classification and organization, where images are treated as documents and local level pixel patterns or other exchangeable visual features

are treated as words (Fei-Fei and Perona 2005; Sivic, Russell et al. 2008; Li, Wang et al. 2010). Second, LDA is easily embeddable within other automated analysis and machine learning systems (i.e. the topic or document topic proportions could then be used as features in SVM or similar methods). Third, the topic mixture model is powerful compared to many other automated discovery systems which simply report a simple match or non-match to a group (here, a fine-grained topical "distance" between two entities can be calculated). This in turn lends itself well to many kinds of visualizations and representations of entities based on the relative strength of correlation between entity and topics, as well as numerous opportunities to use this data to identify networks of relations between entities and topics, entities and entities, and between the individual exchangeable properties. Further, extensions of this methodology are relatively easy to develop and can extend LDA to discover correlations between topics, how collections change over time, and supervised topic analysis, where the power of the Bayesian statistical model can be used to identify known groupings and then used as a fine grain predictor of mixtures over known taxonomies.

LDA provides many opportunities to discover and model latent information and networks beyond the simple relational networks that are a common feature of modern virtual spaces. The technique retains this ability even for social data that do not implement relational functionality. In the subsequent section, we begin to outline and explore some of the potential opportunities available to model virtual organizations via LDA

4. Other LDA Techniques

The application of LDA techniques to help identify collective discursive behavior is an important and useful methodology to better understand topical organization in large sets of social data. Over the last few years, numerous flavors, reconfigurations, and embedded architectures have been developed around LDA. We considered each of these and their potential to study latent relationships and behavior within social media, social networking sites, and virtual organizations. Each of these techniques had their own best fits and limitations.

4.1 Supervised LDA

Supervised LDA (sLDA) is a form of LDA that can be entrained to a specific set of topics (Blei and McAuliffe 2010). In this methodology, each document is additionally associated with a value or word to indicate its group or value. The algorithm works largely the same as with traditional LDA, except it takes into the account this value while maximizing likelihoods. Topics end up maximally distributed between the range of input values. It sets the probability distribution based on the assigned values as opposed to regular LDA, which generatively determines some ideal ways of distributing probability in the case the algorithm uses generative analysis to determine the best distribution to account for the evidence. SLDA models can be used as an LDA-based classifier. Once trained with a body of data, the model can be used to predict the class or value of future, previously unseen documents. One example in our project is the case where one's research may have developed and applied some metric for the measure of some attribute of a message (Blei and McAuliffe 2010). The classic example is training on 1-5 star movie reviews and then guessing star rating based on the text of the review (Ye, Zhang et al. 2009). Another example of such a metric might be an evaluation of the trust shared between users. A hypothesis is that the way this trust is communicated is through the language used with one another (this is a fair assumption when studying social media in which all communication is written). This data set can be paired with the documents to develop topic models, which best predict the observed differences in trust shared between users. This model could then be used to potentially predict trust shown in future previously unseen documents, based on the developed training set. It is important to note that for such predictions to be seen as useful these models would have to be cross validated on the known data as well as testing and evaluating predictions for accuracy.

4.2 Hierarchical LDA

Hierarchical LDA (hLDA) is a version of LDA that models a collection of documents into a latent hierarchy of topics (Griffiths, Jordan et al. 2004). In other words, the model provides a full hierarchy for the latent topics. This means that the latent topics discovered in the collection will also be modeled to fit a hierarchy where child topics are considered subtopics of their broader parent topics. In our case, this

latent hierarchy could be developed and compared to the site-defined hierarchy of discussions forums. This also provides another alternative to cLDA for seeing connections between topics. Latent taxonomies may be useful to understand how language is used within topics and whether there are identifiable subcategories of discourse within each topic.

5. Modeling Virtual Organizations with LDA

The assumption behind our research hypothesis is that opportunities exist, beyond standard quantitative approaches, to model and understand the behavior and discursive operation of a virtual organization (VO). As such, LDA provides a hybrid approach to exploring this behavior. LDA is itself a complex computational technique that generates a large amount of quantitative data on the applied corpus. But this data is often best understood when the approach is guided and informed by qualitative assessments of the results, as can be seen in recent work on Twitter, LDA, and sentiment analysis (Maas, Daly et al. 2011). Specifically, qualitative hypotheses can be rapidly tested with the data generated through the application of LDA techniques with a corpus derived from social technologies. Some of these mixed-method opportunities are discussed in the following section.

5.1 Message-based configuration

A standard base configuration for using LDA to model social data is to take public messages and communication as documents (Maas, Daly et al. 2011). The models can be evaluated at different scales. Taking messages from an entire social media or social networking platform will yield a topic model that reflects the latent semantic groupings of the entire site, a goal of SSN-LDA (Simple Social Network LDA) (Zhang, Baojun et al. 2007). It may also be of interest to develop topic models from individual sections like individual forums where there is already an assumption about a specific topic. Here, LDA-based topic models can allow an even greater, fine-grained understanding of the topical groupings that defined communication in this partition. Carrying this out across multiple partitions allows researchers to develop their own hierarchical topic models where the base hierarchies are researcher defined and leaf groupings are the discovered latent semantic groupings derived from LDA. In the standard configuration,

the topic modeling can be seen as reflecting the latent topics of discussion within a site. It is of interest because it is independent of any platform-defined categories (e.g. threads on forums or blogs). It is also independent of group or shared interests that users might report membership in. The most accurate semantic patterns discovered reflect the patterns of language used throughout the entire virtual space and evaluation of these patterns may lead investigators to recognize latent, hidden patterns and similarity in use of language that is much different than the rigid organizational structures and topical intention suggested by architects and managers of the virtual space. For instance, a social networking site may be developed and intended for use as a discussion group for physicists to discuss topics related to their profession, but a LDA-based topic model might reveal that discussion tend to focus or evolve into a discussion of favorite films, mentors, and books without regard for the intended usage of various sub-forums on the site.

This perhaps indicates that people view the forum as possible home bases which reflect their identity, an identity which they ultimately seek to perform in the social group regardless off the actually chosen forum topic. This research into the level to activity in forums is a performance of identity or subject matter is a sociologically interesting and potentially rich subject for further investigation.

5.2 Author-based configuration

The author-based configuration suggested by Pantel and Penachiotti (2006) combines all discovered posts by a certain author into one document. The corpus then becomes the collection of all user-centric documents for all users of an SNS. In this configuration, the topic model fit to the data becomes like a set of attributes for each user classifying the user based on what topics they choose to post about. Here, the per-document topic proportions can be used to calculate and visualize the similarities between various users. This data can also be used to suggest the existence of various classes of users. The ways in which common types of language are used to discuss similar topics on the site can emerge. This data is independent of user-defined groupings, friends, or common group membership. Natural language analysis can be used to determine whether similar user types tend to communicate with each other or whether

these grouping define forum archetypes, where each sub-forum naturally attracts certain proportions of various user classes.

5.3 Thread-based configuration

In a thread-based configuration, all posts in a given thread would be combined to create thread-level documents. When fitted to the model, topics would likely represent latent similarities between various threads on a site. This configuration would provide results similar to the message-based configuration, except that the generated data would speak more to themes within the body of thread level discussions across the site as opposed to individual posts.

6. Generated Data

6.1 Topic models

The main result generated by the LDA process is the topic model on the body of documents. Here, the topic represents a proposed yet unobserved set of relations and co-occurrence of properties. Each topic model is comprised of a full set of probabilities for each term in the corpus. This probability vector is drawn from the space of all possible probability vectors for terms and represents the possible probability vectors for terms most commonly discovered within the corpus and which best account for all of the observed terms. These possible vectors are discovered generatively. By itself, the topic model represents for each topic the likelihood of the occurrence of each term. By examining the most probable terms for each topic, one can qualitatively hypothesize about what aspects of the text documents an individual topic is reacting to. For example, a topic may be discovered to contain jargon and terminology from a specific domain with high probability. The yet un-named latent topic can be provisionally understood as relating to this domain. Topics can also sometimes be discovered based on language usage. For example, most emails contain some common opening and closing structure (e.g. Dear X, Sincerely yours, Best wishes, and Regards). These structures are often like templates, but the exact language, grammar or placement used in each may be different. Through the transitivity of different permutations around a common

document structure of a topic might be identified, many documents may be identified as containing this topic to some degree.

6.2 Word-to-Topic distribution

Performing an LDA analysis will also generate data by which individual words can be mapped to the discovered latent topics. Here, a probability vector can be derived for each term indicating the probability that the appearance of a word is related to each of the possible topics. Thus, for each term one can identify the topic that uses the word most often. This data helps develop the document to property distribution, but could also allow one to visualize which words in a document are related to which topics by coloring each word based on its most likely topic.

6.3 Document-to-Topic distributions

As previously discussed, once an LDA model has been fit to a given corpus, the model can then be used to calculate estimates of topic proportion from previously unseen documents. These topic vectors can also be visualized. These topic proportion vectors can be used to compute the similarity between documents using the Hellinger Distance formula, a 'symmetric distance between distributions' (Blei and Lafferty 2007). By setting a threshold distance, one could create a network of all documents in a corpus based on their topical similarity. This is especially interesting when using an author-based document configuration for the LDA model as this would create a network of users based on similarity of topics used and language discussed.

7. Methods for Modeling Topics in Social Platforms with Message-based LDA

Because of the wide variety of topics being discussed in forums, blog posts, and profiles, a fundamental research objective was to obtain representations of semantic knowledge. In this section, we discuss our results derived from the application of LDA to the textual corpora of two different, but popular virtual scientific communities of practice which employ social media and social networking technologies. In addition to revealing the utility of LDA to this 'Big Data' project, we highlight some of the challenges

associated with the project as well. In order to run LDA analysis on social data, one must have or 'prepare' a body of documents to be analyzed for latent topic categories. In our case, both life science platforms had online forums. We considered each message in a forum to be a document. We studied data from two different sites developed for use in the life sciences (labeled Site A and Site B). The following sub-sections describe what data from each site was selected for analysis, how the data was prepared, and what parameters were chosen in order to build the models. Support vector machine (SVM) was experimented with, but we quickly found, like others (Sujitha, Selvi et al. 2014), that LDA outperforms the term-based SVM model and topic-based SVM model significantly for our specific corpora.

7.1 Sample Size

Site A is a very large, active site featuring a wide array of discussions forums. Previous qualitative research was conducted by the authors of this paper (Anonymized) where we similarly explored the types of discussions, topics, and trust expressed within these forums and among active users of the site across forums. But evaluating all posts on this site by hand was impractical, expensive, and excessively time consuming. However, we did undertake human investigation of 50 active topics to get a qualitative feel of the corpora. This was one impetus to explore other computational and quantitative methods to derive similar information. Site A is organized as a collection of forums, each of which contains many discussions topics or threads (born by an initial opening post). Any user can create a forum or topic within an existing forum. At the time we were collecting data, Site A contained 890 unique discussion forums (See Figure 2 for a breakdown). A cursory evaluation of this list revealed that a large portion of the forums were largely unused. In fact 278 of these forums contained no topics so there were no discussions to analyze in these instances. Another 262 forums contained one or more topics, but did not contain any topics with replies: again this leaves little to evaluate. In order to discern useful information, we needed to limit our scope to the forums that contained rich discussions for analysis. We therefore developed a methodology to define an active forum and to consider only posts from amongst these forums for analysis. An active forum was defined as a forum that contained at least 5 topics that had received at least 3 responses each. Using this definition, there were 70 active topics on Site A.

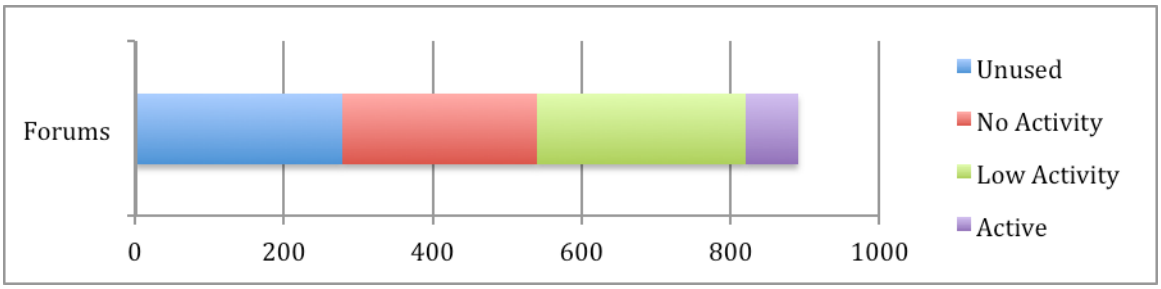

Figure 2: Distribution of forums on Site A by activity

In order to look at consistent sets of data and in the event that the topic and response pattern is indicative of some class of information, we selected 12 of these 70 forums, but sampled evenly throughout the range of active forums. Active forums were grouped into bins based on the amount of activity and one forum was randomly selected from each bin (see Table 1).

Forum	Active Topics	Replies per Topic	Total Posts
Scientist Musicians	5	9.2	72
Italian User Group	10	8.1	114
New User Group	15	14.3333	278
Population Genetics	20	15.25	290
UK Science Policy	26	6.0385	542
Blogging Conference	32	16.6875	580
Women and Science	36	11.9444	395
Bioinformatics	40	7.6	429
News and Opinion	49	19.8571	884
Ask the Editors	51	8.1765	567
Bloggers	56	16.1964	986
Protocol Discussion	62	6.2097	619

Table 1: Breakdown of forums from Site A selected for analysis

With Site B, we had access to the site's complete database, which included all posts on the site to date. Site B also features discussion forums for life scientists in much the same way as Site A. There are forums that contain topics, which constitute threads of discussion. Site B has a higher-level hierarchy,

which categorizes forums into broad sections of site discussion, but otherwise the structure is almost identical. Because we were able to obtain a full set of the discussions on this site, we decided to develop a topic model covering all the data. This was done in part to see the topic model that would result from an unfiltered site, but it also serves as a model of comparison to the results from Site A to evaluate how selection of data influence a topic model. The data, which we analyzed from Site B, consisted of 38,222 posts from 518 forums. Using the same methodology for defining active topics as previously used, Site B contains 139 active topics, and 104 combined topics with no posting activity (see Figure 3). As a fraction of total forums, Site B is much more likely to have forums with some level of posting activity than Site A, and also contains a higher percentage and greater total number of active forums. This difference in composition could have some effect on the topic models derived from each site.

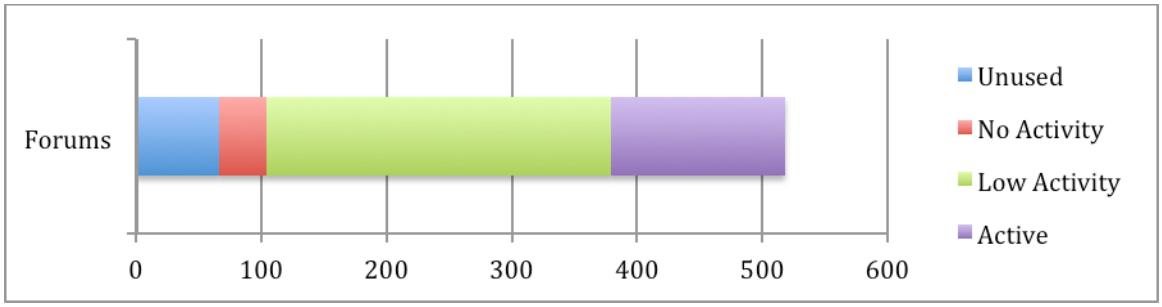

Figure 3: Distribution of forums on Site B by activity

7.2 Data Preparation

The literature on topic models suggests that they benefit from preprocessing the text of a document (Mierswa, Wurst et al. 2006). Among the most common considerations for preprocessing methods are word stemming, common and uncommon word removal, identifying specialized text entities, and the removal of non-important or non-word entities. Each method has its own pros, cons, and effect on the final results in large or subtle ways. Word stemming is a technique of processing a vocabulary of words to reduce its size by reducing all words to their root form. This process eliminates the consideration of plurals and various verb conjugations as unique words. This helps each word form carry maximum meaning (theoretically). The assumption is that keeping all of the stems of a rootable word does not contribute meaningfully when assigning topics. For example, Blei and Lafferty (2006) recommend word

stemming, though they warn that stemming software should be carefully chosen as some are overly aggressive. We agree that careful consideration of the corpus should be done before implementing a word-stemming algorithm prior to LDA analysis and we did not perform word stemming as a pre-processing phase.

Another data preparation issue is the organization of stop words. Some suggest removing "stop" words prior to performing analysis (Blei, Ng et al. 2003). Stop words are considered to be common words, which carry little semantic information. In terms of parts of speech, most articles, pronouns, prepositions, and conjunctions would be considered stop words. Very common nouns and verbs may also be considered stop words, but careful consideration should be given before any word is chosen to be removed from a corpus of documents under consideration. There are many lists of stop words available. They can range in size from dozens to thousands of words. To prepare data for our project, we used a standard list of 311 stop words. In early tests, LDA-based topics models were fit to data collected from both sites without removing any stop words. Stop words dominated the topic model. Another common methodology used to handle stop words is to not remove them prior to fitting the model, but to ignore them in the final topic model.

Depending on the source of a body of text, it may contain various structures and entities which one may wish to remove or encapsulate. We refer to this process as text cleaning. The most common text cleaning operations are to remove punctuation and numbers. But one must determine whether there are any entities within the corpus that contain punctuation and/or numbers which should be preserved as they may carry important information from the dataset. Among these types of entities may be links, email addresses, emoticons, and phone numbers. Depending on your pre-evaluating of the text corpora, a user may decide to preserve these entities or remove them completely. This is an important process because it would be undesirable to simply remove punctuation as this may turn entities like links into a collection of words which carry little or no meaning to the analysis: words like www, com, org, etc. Our collected data sets for Site A and Site B contain HTML tags, so simply removing punctuation might result in adding a lot of words representing HTML tags to the documents like div, span, href, etc. If one wishes to remove

HTML and preserve references to links and email addresses, careful pre-processing must be done prior to analysis to prevent garbage words from entering the vocabulary.

Taking all of these factors into consideration, the following preprocessing steps were conducted on the forum communications collected from Site A and Site B prior to fitting and LDA model to the respective corpora.

1. Convert document to UTF-8.
2. Correct broken HTML
3. Remove all tags except links
4. Convert links to entities representing only the link to domain but not individual pages
5. Convert email addresses to entities
6. Convert statements of form <letter>& <letter> to entities as in B&B(Bed and Breakfast) or R&D (Research and Development)
7. Identify and preserve 28 kinds of common emoticons representing smiles, frowns and hearts.
8. Remove punctuation and HTML entities, taking into consideration various character encodings. Exception for dashes and underscores.
9. Remove stop words
10. Remove repeated dashes and orphan dashes and underscores (not attached to a word)
11. Remove extra spaces.
12. Convert to lowercase.
13. Generate a vocabulary from the unique words remaining.
14. Take each document to be a spare vector on the vocabulary containing the count of each of the unique word in a given document.

The final vocabulary contained 32,767 words drawn from 5,703 documents for Site A and 95,803 words from 36,576 documents from Site B. This number is larger than it needs to be because of the inclusion of domain links.

7.3 Models Generated and LDA Parameter Tuning

Following Hazen (2010), models were generated to discover 40 latent topics within the data from Site A and Site B. In LDA, a good fit is one where the latent topics cover the actual range of different word occurrences. LDA allows one to suggest an original α and we set α to 1 (which causes a Dirichlet to output uniform distribution). The original β was randomly assigned. α and β were iteratively optimized to decrease the divergence of the estimates from the evidence (corpus). As we did not have an estimate of the mixing level, we did not provide a different α to 1. We used Blei et al.'s (2003) original code which implements the variational inference algorithm rather than Gibbs sampling.

8. Results

8.1 Site A Results

The data generated for each latent topic is a vector containing the probability of every word in the vocabulary. A common way to display a topic is to only list the highest probability words in each topic. For most topics, this is sufficient because often there are only a handful of words that are strongly indicative of belonging to a certain topic, with a steep drop-off in probability after that with the vast majority of words in the vocabulary having near nil probability for the topic. Table 2 illustrates the top ten most probable words for the first 20 topics of the topic model generated by from the sample of documents from site A.

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
opinion, surname, find, name, science, article, people, comment, see, married	science, nature, protocols, scientific, religion, scientists, system, selection, natural, research	science, people, space, libel, scientists, time, society, legal, see, public	research, work, people, years, sharing, time, data, science, take, two	drugs, cognitive, people, drug, authors, enhancement, taking, brain, enhancers, effects
Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
science, sessions, session, di, ideas, day, scientific, conference, unconference, animals	research, scientific, government, science, policy, public, advice, consultation, federal, funding	data, university, research, usa, biological, mining, new, bioinformatics, papers, students	motif, evolutionary, print, file, russian, found, struggle, todes, new, theory	blog, blogging, science, conference, blogs, people, network, posts, bloggers, nature
Topic 10	Topic 11	Topic 12	Topic 13	Topic 14
science, research, cfse, mr, obama, first, committee, new, select, education	people, time, technorati, seems, google, post, culture, determined, article, work	research, data, people, misconduct, scientific, comment, studies, ref, new, barbie	d, gst, differentiation, fst, population, populations, alleles, allele, measure, diversity	perl, bioinformatics, work, program, scientists, assay, interesting, people, help, see
Topic 15	Topic 16	Topic 17	Topic 18	Topic 19

free, science, determined, human, freedom, behaviour, people, question, decisions, random	intelligence, iq, research, differences, race, science, genetic, human, people, rose	human, genetic, diversity, possibility, intelligence, found, people, group, does, article	vector, gel, dna, cells, using, pcr, ligation, u, tried, plasmid	research, science, death, uk, two, funding, data, evidence, work, made

Table 2: Top ten most probable words for 20 topics of the topic model generated from Site A

Even from looking at a small fraction of the words generated for each topic, one quickly gets a sense of the topics discussed within Site A (and they are highly correlated with findings from qualitative work we have done). A cursory inspection might place these latent topics as beginning to agree with some of the known forum discussions (see Table 3). Of course, there is no one-to-one mapping. Rather, each document contains a mixture of words from perhaps many topics.

Population Genetics	1,13,16,17
UK Science Policy	2,6,10,19
Blogging Conference	5,9
Women and Science	0,12
Bioinformatics	7,14
News and Opinion	0,2,8,11,12,
Bloggers	9
Protocol Discussion	1,18

Table 3: Some possible observed correlations between observed topics and known discussion forums

Also of interest are topics like topic 40 which contains a list of pleasantries and words common in all forum communications like: good, idea, looking, see, great, find, :), say, best, and link. All documents likely contain some proportion of words from this topic.

8.2 Site B Results

The results from Site B are interesting because they represent the latent topics from an entire virtual scientific platform. The data includes information from every post in every category on the site. This

provides an opportunity to understand the types of communications taking place on the site without having to rely on the site's own reported hierarchy of topics and threads. Many of these discussion boards are used very little so in terms of the site-reported hierarchy, they represent topics which are pre-approved by the site's administration, but the LDA-based topic model reveals the topics people are using the site to discuss and which they are not. Table 4 illustrates the top ten words from the first 20 topics of the model.

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
research, experience, post, work, date, centre, project, closing, university, biology	chemistry, chemical, materials, organic, engineering, university, structure, surface, synthesis, developed	j, e_frown, s, m, p, k, r, c, morpholino, h	science, new, scientists, research, world, technology, said, scientific, life, public	nobel, university, prize, work, first, professor, years, born, chemistry, research
Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
sample, method, standard, column, samples, using, water, analysis, phase, concentration	ml, add, solution, cells, stock, plate, medium, volume, c, tube	image, fluorescence, light, imaging, images, microscope, using, cells, fluorescent, microscopy	rna, kit, tissue, protocol, dna, extraction, used, using, isolation, method	protein, gel, proteins, buffer, membrane, sample, gels, using, bands, run
Topic 10	Topic 11	Topic 12	Topic 13	Topic 14
proteins, protein, molecules, amino, two, acid, structure, molecule, acids, form	glycine, non-polar, alanine, serine, cysteine, aspartic, acid, acidic, lysine	find, site, link, information, forum, help, post, web, good, search, www.siteb.com*	solution, water, staining, pbs, acid, minutes, wash, sections, sodium, tissue	biology, molecular, cell, research, phd, techniques, project, biochemistry, cellular, position
Topic 15	Topic 16	Topic 17	Topic 18	Topic 19
problem, see, instrument, time, issue, system, check, pump, plasma, new	good, time, lab, used, work, using, best, better, different, people	cell, current, solution, patch, channels, pipette, channel, using, potential, seal	cancer, clinical, disease, research, drug, therapy, medical, patients, treatment, months	drug, activity, collagen, assay, protein, inhibitor, receptors, receptor, binding, kinase

Table 4: Top ten most probable words for the first 20 topics of the topic model generated from Site B. *(* This term reference links to pages on site B. the name of the site is anonymized*)

The topics in Table 4 reveal similarities and marked differences from the model for Site A. First, the topics are highly technical. Site B is known to be primarily focused on the sharing and discussion of

scientific protocols and procedures relating to various research methods in the life sciences, whereas Site A does not actively promote such discussions. Any similar topics only emerge organically. In Table 4, Topic 6 on procedures involved mixing solutions and volumes; Topic 7 on fluorescence microscopy; and Topic 8 on procedures for working with and isolating DNA. Other topics of note are Topic 2, which appears to be related to citations (the single letters were determined to be people's initials). Topic 4 appears to involve language for discussing biography or credentials. Topic 12 revolves around language directing people to information, often involving links to other pages within the site. A reasonably clear delineation between topics, which discuss a certain field of research within the life science and those that do not emerges. This is expected, because each scientific domain has its own vocabulary and argon separate from other types of language like that related to papers, or job postings. These distinctions seem to have been well delineated by the topic model in LDA natural language processing.

9. Correlated LDA

We also applied a similar topic model algorithm known as correlated LDA (CLDA), which creates a topic model that is different from standard LDA in the way it samples documents and records the per-document likelihood (Blei and Lafferty 2007). The way the data is structured allows another companion algorithm to compute relationships between topics within a user-defined threshold. The connections between topics represent instances where words from one topic occur that are correlated with the appearance and use of words from the second topic. Two topics can be related when they often occur together with significant proportion within observed documents.

9.1 Methodology

The methods used to prepare for the CLDA application are identical to those of standard LDA. We used the same methods described in the previous section to develop a correlated topic model. The algorithm that reports correlations between discovered topics involves setting a user-defined threshold. This defines the level of relationship between topics necessary for them to be considered related. The threshold can be between 1 and 0, with the latter resulting in no correlation reported and 1 reporting a fully connected topic

graph. The algorithm determines a value between 0 and 1, denoting the strength with which each pair of topics is correlated. This threshold serves as a cutoff to only show the pair that is more correlated than the chosen value. For both Site A and Site B, the threshold chosen was 0.33 after sampling a variety of thresholds to not over or under fit the data.

9.2 Results

The results of the cLDA analysis reveal how applying LDA techniques to a large social dataset provides important insights into the overall sets of topics across traditional virtual organizational structures (such as threads). These initial results indicate some known relationships between various types of discussion on each site. It also highlights potential concerns in understanding how to set analysis parameters. Most importantly, it also suggests relationships that were not known or observed prior to the analysis. These types of results would likely play the most important role in understanding the latent discourse in the virtual organizations we studied. More broadly, CLDA is particularly useful in discerning sets of topics in large online corpora.

9.3 Site A Results

A map of the relationships between discovered topics is shown in Figure 3. The topic model is not highly correlated. Of all the pairs of correlated topics meeting the threshold criteria, only 26/40 have any determined relationship with another topic. Also, there are 4 separate graphs that do not interact with each other. The main group is also weakly connected to the rest of the topic networks.

Figure 4: Map of correlated topics discovered in Site A

Evaluating the large main group in the center of Figure 4, there appears to be two highly connected topics. They both relate to blogging, the site, and keywords like "science." Several of the sampled forums were about science blogging so this derived topic list is robust. The blogger forum has a large number of posts about blogging in general and a specific forum existed to organize a conference for science bloggers. These two categories are highly correlated, and both cover topics about conferences.

This main group also has a weak connection via a perl programming and bioinformatics topic to a group that is interested in cell and protein research. This group is weakly connected to a linear group of other topics. Furthermore, there is a correlation group that represents different topics related to the science policy forum. Two additional stranded pairs seem to correlate discussion of cognitive enhancement drugs to bioinformatics and a pair of topics both of which seems to contain different terminology within the field on population genetics, which was one of the sampled forums. This is in congruence with our earlier, published qualitative work done on Site A (Anonymized).

These results signal that many of the sampled forums may have one or two distinct latent topics which do not overlap greatly with the other sampled forums. Many of the omitted topics probably contain highly forum-specific topics which contain terms which are not used within other topics or they could be topics which represent language which is common across all topics, but not highly correlated with the topics. We would expect the type of topics discovered that cover general language like the pleasantries topic mentioned previously. A topic model drawing from a more robust sample, yet still limited to 40 topics might yield a more connected correlation chart as topic will be forced to combine language which in this small topic model can be represented as separate yet correlated topics.

9.4 Site B Results

The map of correlated topics from Site B (see Figure 5) is more connected than that of Site A. This is despite only connecting an identical 26/40 topics. In Site B, there is only one disconnected topic pair, which appears to be about application for either job or research opportunities. The other topics are more highly meshed.

Figure 5: Map of correlated topics discovered in Site B

Also, the dynamics are different here. In Site B, the central node is about information with various topics connecting off from the main hub and associating with each other by individual topics. A comparison indicates that some of the topics we witnessed previously about very specific life science protocols are either not included in the map or are only weakly linked with other topics. This may indicate that the domains of science discussed on this site are so specific that there is little or no high level correlation between those domain-based topics. The most connected topic (degree of 6) covers databases and information and the second most connected topic (degree of 5) covers topics of publishing in open access. Importantly, the topic categories being considered for publication emerge as well (gene sequencing and cancer are prominent). In this specific case, LDA delivered robust results which were in congruence with our extensive qualitative work, which included interviewing and ethnographic participant observation.

10. Conclusion

NLP is not a static object of study. Cambria and White (2014) usefully speculate on the future of NLP based on 50-year eras. They argue that we are currently in an NLP of a 'Bag-of-Words' and it is a roughly 100 year curve to move through 'Bag-of-Concepts' to 'Bag-of-Narratives'. This article has argued that though we have not crossed into the 'Bag-of-Concepts' era, NLP continues to evolve. One of its evolutions has been the application of NLP outside of computer science and, in our case, in an interdisciplinary project in the social sciences, which sought to render visible topics and trends across a 'Big Data' scale of text. NLP has tremendous promise across many disciplines. Without the implementation of natural language processing techniques (in our case, Latent Dirichlet Allocation (LDA)), it would not have been possible to understand and decipher some of the nuanced conversations that were occurring across a multitude of threads. Specifically, topics within these platforms and discussion threads and other social media spaces did not always accurately encapsulate the topics of conversation within the meta-thread. Additionally, topics, conversations, and discussions often digressed like in face-to-face conversation. By using natural language processing, we were able to decipher key topics by particular groupings of individuals or forums. This allowed us to get a much better understanding of the ways in which these scientific virtual organizations collaborated, innovated, and created knowledge. This has immense utility not only within the organizational context, but within natural language processing as an interdisciplinary field.

As Blei (2012) notes, topics and topical decompositions are not in a sense 'definitive.' Fitting a model to any collection will yield patterns regardless of whether they exist in a true sense the corpus. But that is where LDA pairs so well with qualitative techniques. This presents new ways for natural language processing to assist and guide qualitative research and research design. In this case, LDA can help identify areas for further explorations as well suggest a variety of different possible groupings the combinations of which can help to best understand the nature and activities of contemporary social media, social networks, and virtual organizations – all of which have large data footprints. Another important

contribution of this article is that we found that LDA can move us beyond the state-of-the-art in conventional sociological techniques; in this case, Social Network Analysis techniques.

Acknowledgments

The authors wish to thank Rebecca Fiebrink for her comments on an earlier version of this paper. This material is based in part on work supported by the National Science Foundation under Grant no. 1025428. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Blei, D. M. (2012). "Probabilistic topic models." Commun. ACM **55**(4): 77-84.

Blei, D. M. and J. D. Lafferty (2006). Dynamic topic models. Proceedings of the 23rd international conference on Machine learning. Pittsburgh, Pennsylvania, ACM**:** 113-120.

Blei, D. M. and J. D. Lafferty (2007). "A correlated topic model of science." The Annals of Applied Statistics: 17-35.

Blei, D. M. and J. D. McAuliffe (2010). "Supervised topic models." arXiv preprint arXiv:1003.0783.

Blei, D. M., A. Y. Ng, et al. (2003). "Latent dirichlet allocation." The Journal of machine Learning research **3**: 993-1022.

boyd, d. m. and N. B. Ellison (2007). "Social Network Sites: Definition, History, and Scholarship." Journal of Computer-Mediated Communication **13**(1): 210-230.

Brzozowski, M. J. (2009). WaterCooler: exploring an organization through enterprise social media. Proceedings of the ACM 2009 international conference on Supporting group work. Sanibel Island, Florida, USA, ACM**:** 219-228.

Buntine, W. L. (1994). "Operations for learning with graphical models." Journal of Artificial Intelligence Research **2**(1): 159-225.

Cambria, E., B. Schuller, et al. (2013). "New Avenues in Opinion Mining and Sentiment Analysis." Intelligent Systems, IEEE **28**(2): 15-21.

Cambria, E. and B. White (2014). "Jumping NLP curves: A review of natural language processing research." IEEE Computational Intelligence Magazine **9**(2): 48-57.

Fabisch, A., Y. Kassahun, et al. (2013). "Learning in compressed space." Neural Networks **42**(0): 83-93.

Fei-Fei, L. and P. Perona (2005). A bayesian hierarchical model for learning natural scene categories. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, IEEE.

Fombrun, C. J. (1982). "Strategies for Network Research in Organizations." Academy of Management Review **7**(2): 280-291.

Geman, S. and D. Geman (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." Pattern Analysis and Machine Intelligence, IEEE Transactions on(6): 721-741.

Geyer, W., C. Dugan, et al. (2008). Recommending topics for self-descriptions in online user profiles. Proceedings of the 2008 ACM conference on Recommender systems. Lausanne, Switzerland, ACM**:** 59-66.

Griffiths, T., M. Jordan, et al. (2004). "Hierarchical topic models and the nested Chinese restaurant process." Advances in neural information processing systems **16**: 106-114.

Gruber, T. (2008). "Collective knowledge systems: Where the Social Web meets the Semantic Web." Web Semantics: Science, Services and Agents on the World Wide Web **6**(1): 4-13.

Hazen, T. J. (2010). Direct and latent modeling techniques for computing spoken document similarity. Spoken Language Technology Workshop (SLT), 2010 IEEE.

Kaisler, S., F. Armour, et al. (2013). Big Data: Issues and Challenges Moving Forward. System Sciences (HICSS), 2013 46th Hawaii International Conference on.

Kantardzic, M. (2011). Data mining: concepts, models, methods, and algorithms, John Wiley & Sons.

Kireyev, K., L. Palen, et al. (2009). Applications of topics models to analysis of disaster-related twitter data. NIPS Workshop on Applications for Topic Models: Text and Beyond.

Li, L.-J., C. Wang, et al. (2010). Building and using a semantivisual image hierarchy. Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE.

Maas, A. L., R. E. Daly, et al. (2011). Learning word vectors for sentiment analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. Portland, Oregon, Association for Computational Linguistics**:** 142-150.

Mann, G. S., D. Mimno, et al. (2006). Bibliometric impact measures leveraging topic analysis. Digital Libraries, 2006. JCDL '06. Proceedings of the 6th ACM/IEEE-CS Joint Conference on.

Mierswa, I., M. Wurst, et al. (2006). YALE: rapid prototyping for complex data mining tasks. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. Philadelphia, PA, USA, ACM**:** 935-940.

Mowshowitz, A. (1997). "Virtual organization." Commun. ACM **40**(9): 30-37.

Pak, A. and P. Paroubek (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. LREC.

Pantel, P. and M. Pennacchiotti (2006). Espresso: leveraging generic patterns for automatically harvesting semantic relations. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Sydney, Australia, Association for Computational Linguistics**:** 113-120.

Pritchard, J. K., M. Stephens, et al. (2000). "Inference of population structure using multilocus genotype data." Genetics **155**(2): 945-959.

Scott, J. and P. J. Carrington (2011). The SAGE handbook of social network analysis, SAGE publications.

Sivic, J., B. C. Russell, et al. (2008). Unsupervised discovery of visual object class hierarchies. Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE.

Sujitha, S., S. Selvi, et al. (2014). "Emotion classification of textual document using Emotion-Topic Model." International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) **3**(2).

Travica, B. (1997). The Design of the Virtual Organization: A Research Model. Americas Conference on Information Systems, Indianapolis.

Wei, X. and W. B. Croft (2006). LDA-based document models for ad-hoc retrieval. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. Seattle, Washington, USA, ACM**:** 178-185.

Wellman, B., A. Q. Haase, et al. (2001). "Does the Internet increase, decrease, or supplement social capital? Social networks, participation, and community commitment." American behavioral scientist **45**(3): 436-455.

Ye, Q., Z. Zhang, et al. (2009). "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches." Expert Systems with Applications **36**(3): 6527-6535.

Zhang, H., Q. Baojun, et al. (2007). An LDA-based Community Structure Discovery Approach for Large-Scale Social Networks. Intelligence and Security Informatics, 2007 IEEE.