RESEARCH STATEMENT

JORDAN BOYD-GRABER

My long term research goal is to create models that can capture the intricacies of language and can interact in changing, uncertain environments. I approach these problems using the techniques of machine learning and probabilistic models, mathematical formalisms for encoding knowledge gleaned from data and making predictions about the unknown.

The probabilistic method has had a great impact on our daily lives. Every e-mail we read has been examined by a probabilistic algorithm to determine whether it is spam or not; Amazon and Netflix tell us what movies and books to read based on our past interests; and SIRI intuits our intent from natural language speech. These techniques have helped us cope with the onslaught of "big data" from document collections and social media, but as successful and popular as these techniques are, their underlying probabilistic models are often inscrutable.

My research focuses on a class of probabilistic algorithms called *topic models* that help information consumers navigate large datasets. Given a large collection of documents, topic models discover the constituent themes in a corpus. For example, given a decade's worth of New York Times articles and the number of topics K you want the model to discover, topic models can discover themes related to "business", "technology", and the "arts" (Figure 1).

These models are popular in both academia, industry, and government. In academia, latent Dirichlet allocation, the prototypical topic model, is the most-cited article on the bibliography sharing service Mendeley. In industry, e-Discovery firms use topic models to discover themes that litigants are withholding in a mountain of data. In government, the NIH uses topic models to monitor their research portfolio and the FDA to monitor tobacco advertisements.

In this document I will discuss techniques that enable machine learning algorithms that respect the human users that machine learning ostensibly serves and can scale to big, real-world datasets. In the rest of this research statement, I discuss

- techniques to **evaluate** the output of topic models,
- techniques for users to **interact** with models in a tight loop,
- scalable techniques to allow these models to discover **new words** in streaming data, and
- **applications** of topic modeling to discovering political slant and framing.

While my research focuses on topic models, the themes in my research connect with broader themes in machine learning:

- building **interpretable** models and
- using **parallel** and **streaming** algorithms that adapt to new data or when data is too big to fit on a single machines;
- together, these lead to **scalable** machine learning algorithms that humans can **use and understand**.

1. Evaluating Topic Models

For a long time, topic modeling was often a "take it or leave it" proposition. You took a corpus, ran topic modeling on it, and then either you liked the output or you didn't.

Statisticians had ways of measuring how good a topic is. You can ask what the probability of a dataset is under a model that you learned from a corpus. The higher this number—called *held-out*

Date: March 14, 2013.

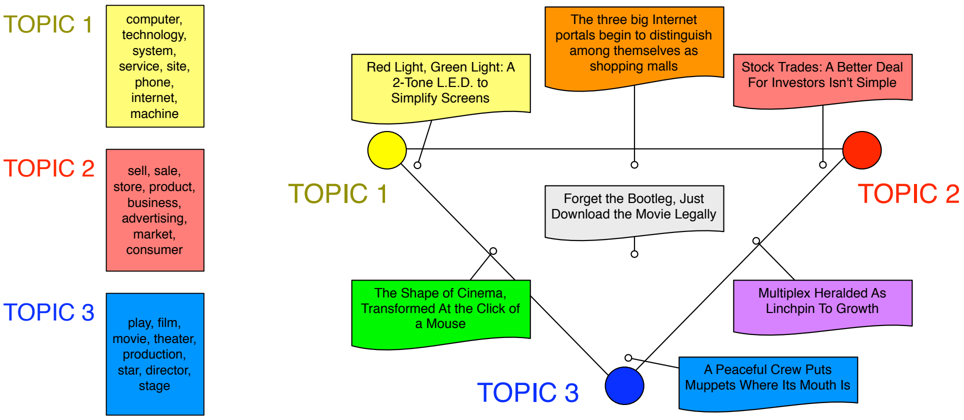
TOPIC 1
computer, technology, system, service, site, phone, internet, machine
TOPIC 2
sell, sale, store, product, business, advertising, market, consumer
TOPIC 3
play, film, movie, theater, production, star, director, stage
Red Light, Green Light: A 2-Tone L.E.D. to Simplify Screens
The three big Internet portals begin to distinguish among themselves as shopping malls
Stock Trades: A Better Deal For Investors Isn't Simple
TOPIC 1
Forget the Bootleg, Just Download the Movie Legally
TOPIC 2
The Shape of Cinema, Transformed At the Click of a Mouse
Multiplex Heralded As Linchpin To Growth
A Peaceful Crew Puts Muppets Where Its Mouth Is
TOPIC 3

Figure 1. Topic models are a broad class of statistical techniques that, given texts as input, produce a set of "topics" that describe the broad themes present in the collection (left). For example, given a collection of New York Times articles, a topic model might discover a "business" topic, a "technology" topic, and an "entertainment" topic, all without any category labeling or annotation applied to the text. Each document is also associated with a distribution over topics (right).

likelihood is, the better your model can represent the corpus. This venerable metric makes sense when your goal is to predict words (for example, when typing a sentence on your phone).

But that is not how topic models are used. They guide users through large corpora. In particular, users focus on topics' **most probable words** (i.e., "computer" in the technology topic in Figure 1); held-out likelihood focuses on the entirety of the distribution. Recognizing this disconnect, my collaborators and I developed a simple test of how **interpretable** a topic is; take a topic and insert an *intruder* word [Chang et al., 2009]. If a user can detect the intruder, your topic makes sense. Otherwise, your topic does not make sense.

Moreover, we showed that the conventional measures of topic model quality are not positively correlated with our proposed interpretability measure. So topic model users should measure what they care about. Since our paper, other researchers have built fully automatic metrics that measure how interpretable a topic is and have extended the approach to phrases (rather than just single words).

2. Building better models the probabilistic method

Now that we know how to evaluate topic models, we can start to build better topic models that **improve how topic models are used**. To discuss this process, I have to give a bit of background on how I approach a research problem.

My first step is to posit a *generative model* that tells a story using the language of probability of how I believe data (such as a collection of documents) came to be. This story often contains missing pieces called *latent variables* that we need to discover using the mathematical tools of posterior inference.

For topic models, the standard model, called latent Dirichlet allocation, assumes that each topic is a distribution over words drawn from a Dirichlet distribution. In addition, each document is also drawn from a Dirichlet distribution over topics. Dirichlet distributions encourage sparsity; a document can be about two or three topics, but not all of them.

By changing the underlying assumptions of the model, topic models can capture richer concepts and be more useful to end-users. Before these techniques, it was nigh impossible for users who weren't machine learning experts to tweak a topic model to reflect prior knowledge, fix disambiguation errors, or satisfy end-user needs. But these users, who know their data better than anyone else, can clearly see that topic models are "going bad".

With my students Yuening Hu and Brianna Satinoff, I developed a procedure interactive topic modeling [Hu et al., 2011] that puts a human in the loop.[1] Users can react to the results of topic modeling and refine the results in an interactive, gradual way that creates models that are still statistically sound but also reflect the intuitions and needs of the end user. Conceptually, these models work by slowly building a representation of how a user's mental lexicon is structured. For example, if a user is interested in a dataset of newspaper articles from 1980–2000, a topic model might cluster documents about the Soviet Union and the Russian Federation into two different topics; with our framework, a user who views both as part of a single narrative could tell the system that "soviet" and "russia" are connected words.

We can accomplish this by changing the underlying model. Instead of assuming that an individual topic comes from a Dirichlet distribution, we posit that topics come from a tree-structured distribution over words. The topic distributions are now correlated based on words that should belong together (e.g., "soviet" and "russia"). Another example is shown in Figure 1.

Topic	Types
318	bladder, sci, spinal_cord, spinal_cord_injury, spinal, urinary, urinary_tract, urothelial,injury, motor, recovery, reflex, cervical, urothelium, functional_recovery

Topic	Types
318	sci, spinal_cord, spinal_cord_injury, spinal, injury, recovery, motor, reflex, urothelial, injured, functional_recovery, plasticity, locomotor, cervical, locomotion

Table 1. One topic from 700 topics extracted by topic model on NIH proposals before (left) adding a negative correlation (between "bladder" and "spinal_cord") and after (right). This was done to reflect end-user preferences (reflecting NIH-internal divisions). After the correlation was added to push urinary system terms (in red) away from the motor nerves terms (in blue), most urinary terms went away (in green), and some new terms related with motor nerves appeared(in green).

Conceptually, this representation of topics grew out of a desire to fuse the insights from lexical semantics models with data-driven topic models [Boyd-Graber et al., 2007]. Later, other researchers extended this formalism to database constraints and first-order logic programs.

The other key technical insight is to escape from local minima using online inference. Feedback from users identifies subsets of the data where the model is wrong. We then forget those data, replace the existing model, and then relearn the model, pretending we are seeing some of the data for the first time.

3. Scalable, Extensible Models

As overused as the term "big data" has become, our data really have become big. Topic models offer a way for organizations to get a sense of what these gargantuan datasets have inside them. One of my primary goals has been to allow topic modeling algorithms to scale effectively. There are two common approaches: throw more machines and the problem and move to streaming algorithms.

We've done both. With my student Ke Zhai, we developed scalable algorithms for topic modeling using variational inference and MapReduce [Zhai et al., 2012]. I've also worked with him to develop algorithms to develop techniques that can process streaming datasets.

One fundamental difference of batch algorithms and online algorithms is that in batch settings we know **all** of the words we'll see. This is at odds with the modeling assumptions typically assumed by topic models. Recall that the topic distribution is assumed to be a multinomial distribution over a **fixed** set of words.

[1] This work was supported by National Science Foundation grant #0705832

Particularly for streaming algorithms, this is neither reasonable nor appealing. There are many reasons immutable vocabularies do not make sense: words are invented ("crowdsourcing"), words cross languages ("Gangnam"), or words common in one context become prominent elsewhere ("vuvuzelas" moving from music to sports in the 2010 World Cup). To be flexible, topic models must be able to capture the addition, invention, and increased prominence of new terms.

Allowing models to expand topics to include additional words requires changing the underlying statistical formalism. We use Bayesian nonparametrics to build a distribution over *every possible* word that can be generated by a distribution over all possible strings. This character n-gram model can be thought of as a monkey at a typewriter writing English-like words which can either be validated or rejected by our model. Figure 2 shows how new words can be incorporated into our model.

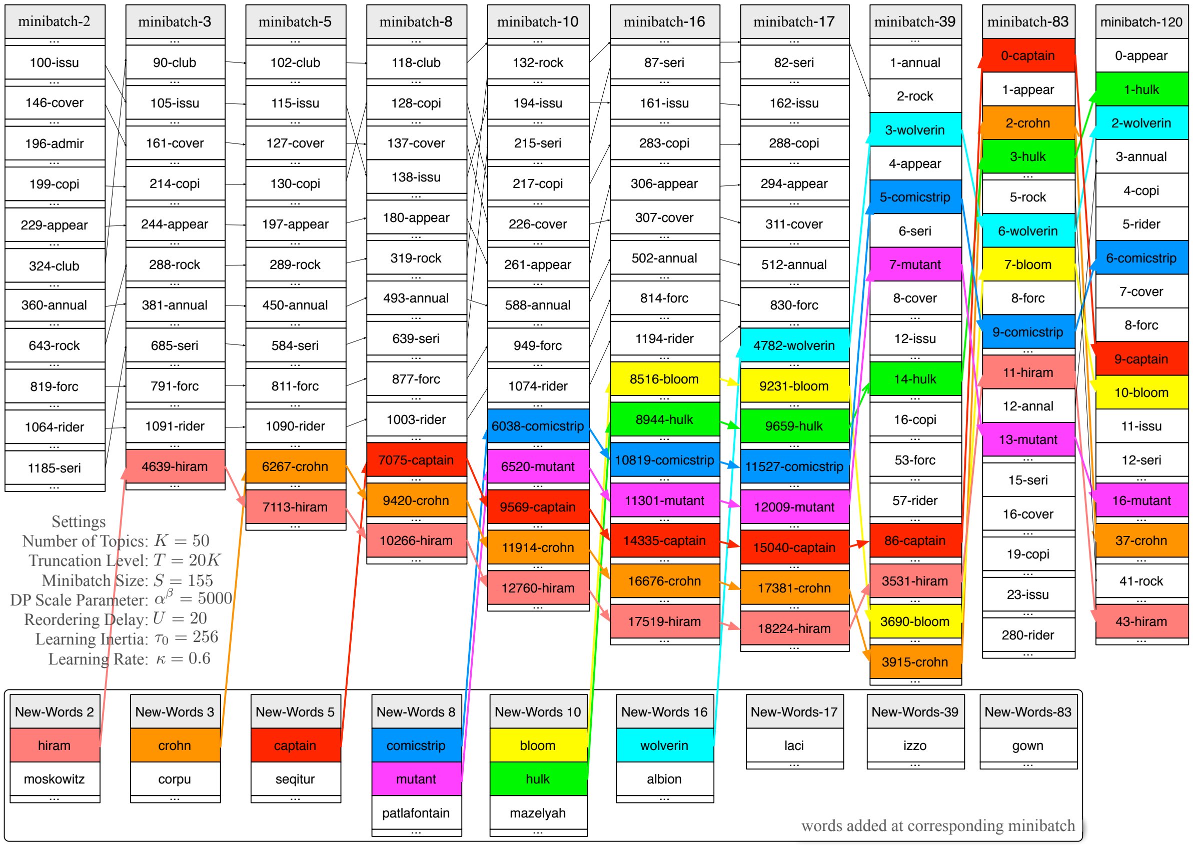

FIGURE 2. The evolution of a single "comic book" topic from an online discussion forum. Each column is a ranked list of word probabilities after processing a subset of documents. The box below the topics contains words seen for the first time in that subset. For example, "hulk" first appeared in the then subset of documents, but became the second most important word by the final minibatch. Traditional models would ignore these new words. Colors help show words' trajectories.

The need for scalable inference goes beyond vanilla topic models; all Bayesian techniques require tools to process large amounts of data. To address these critical problems, we have developed techniques that reduce traditional inference for the transformed Indian buffet process from an $O(n^2)$ operation to a $O(n \log n)$ operation using a fast-Fourier transform, making the technique applicable

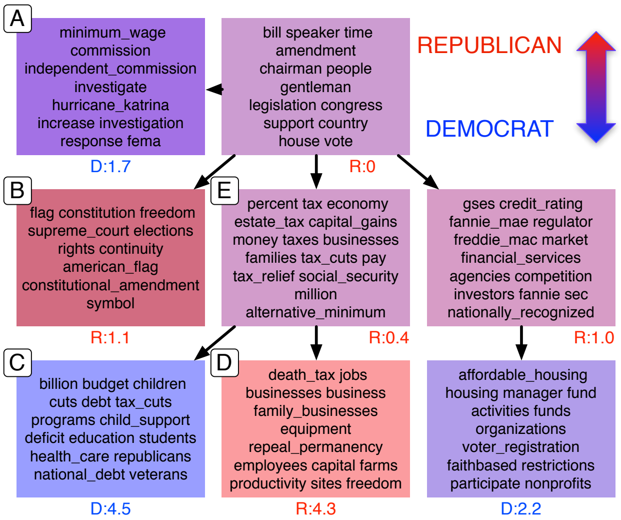
A
minimum_wage commission independent_commission investigate hurricane_katrina increase investigation response fema
D:1.7
bill speaker time amendment chairman people gentleman legislation congress support country house vote
R:0
REPUBLICAN
DEMOCRAT
B
flag constitution freedom supreme_court elections rights continuity american_flag constitutional_amendment symbol
R:1.1
E
percent tax economy estate_tax capital_gains money taxes businesses families tax_cuts pay tax_relief social_security million alternative_minimum
R:0.4
gses credit_rating fannie_mae regulator freddie_mac market financial_services agencies competition investors fannie sec nationally_recognized
R:1.0
C
billion budget children cuts debt tax_cuts programs child_support deficit education students health_care republicans national_debt veterans
D:4.5
D
death_tax jobs businesses business family_businesses equipment repeal_permanency employees capital farms productivity sites freedom
R:4.3
affordable_housing housing manager fund activities funds organizations voter_registration faithbased restrictions participate nonprofits
D:2.2

FIGURE 3. Topics discovered from Congressional floor debates. Many first-level topics are bipartisan (purple), while lower level topics are associated with specific ideologies (Democrats blue, Republicans red). For example, the “tax” topic (E) is bipartisan, but its children have distinct ideological frames. Its Democratic-leaning child (C) focuses on “child_support”, “children”, “education”, and “health_care”, while its Republican-leaning child (D) focuses on the “death_tax”, “jobs”, and “family_businesses”. The number below each topic denotes the magnitude of the learned regression parameter associated with that topic. Colors and the numbers beneath each topic show the regression parameter associated with the topic.

to real-world images for the first time [Hu and Boyd-Graber, 2012] and factorizing the conditional distribution of Gibbs sampling in Dirichlet forests to improve running-time [Hu et al., 2012].[2]

4. Application of Topic Models: Understanding and Interacting with Humans

In addition to core modeling advances, another key component of my research agenda is creating new applications that use topic models and other machine learning techniques to expose how humans use and process information. One specific example of this is allowing computers to detect socially-relevant features from a user's behavior. I discuss this work as reflected in three interconnected problems: sentiment, values, and influence.

Part of this research has included how syntax changes how sentiment is expressed in documents [Sayeed et al., 2012], how sentiment and persuasion is expressed across languages and cultures [Boyd-Graber and Resnik, 2010, Anand et al., 2011], and how individuals can control the topic and influence others [Nguyen et al., 2012]. Recently, we have been using topic models to discover how individuals frame topics in ways that reflect their sentiment or ideology through joint models of what people say and an individual's ideological slant. This discovers, for example, that republicans focus on the costs of taxes (e.g., hurting job creators) while democrats focus on the benefits that taxes provide (Figure 3).[3]

In addition to building models that seek to understand how humans communicate, we are also developing algorithms that can compete with humans. I developed techniques that combine reinforcement learning on incrementally revealed text with hierarchical Bayesian representations of

[2] This research was supported by US National Science Foundation Grant NSF grant #1018625.

[3] This research was funded in part by the Army Research Laboratory through ARL Cooperative Agreement W911NF-09-2-0072 and by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory.

knowledge. This technique improves state-of-the-art batch classification algorithms and can beat humans in a trivia competition [Boyd-Graber et al., 2012].

5. Future Research

My future research will explore generalizations and applications of the work described above.

- **Machine Translation** Recently, we showed that topic models can improve state-of-the-art machine translation algorithms [Eidelman et al., 2012], and we hope to expand those findings to interactive and multilingual models topic models.
- **Generalizing Streaming to Nonparametric Grammars** Our streaming infinite-vocabulary uses a technique for inference that could be generalized to a much broader class of models that could be used for grammar induction, perspective analysis [Hardisty et al., 2010], and "sticky" topic models.
- **Adding Linguistic Rigor** While previous topic models have combined topic models with syntax [Boyd-Graber and Blei, 2008] and morphology [Boyd-Graber and Blei, 2009], these have not achieved state of the art performance on standard linguistic test sets. Combining these models with more sophisticated linguistic theories could improve core NLP tasks and enable new applications such as cross-dialect topic modeling.
- **Assistive Devices** Data from large corpora have been used to improve assistive devices for individuals with language impairment, both by commercial vendors—Lingraphicare—and by our research group as a part of broader effort to improve technology for individuals with motor, language, and vision impairment [Boyd-Graber et al., 2006, Ma et al., 2009]. I will continue to explore these applications to build information frameworks that can assist those with cognitive impairments.

References

[Anand et al., 2011] Anand, P., King, J., Boyd-Graber, J., Wagner, E., Martell, C., Oard, D. W., and Resnik, P. (2011). Believe me: We can do this! In *The AAAI 2011 workshop on Computational Models of Natural Argument.*

[Boyd-Graber and Blei, 2008] Boyd-Graber, J. and Blei, D. M. (2008). Syntactic topic models. In *Proceedings of Advances in Neural Information Processing Systems.*

[Boyd-Graber and Blei, 2009] Boyd-Graber, J. and Blei, D. M. (2009). Multilingual topic models for unaligned text. In *Proceedings of Uncertainty in Artificial Intelligence.*

[Boyd-Graber et al., 2007] Boyd-Graber, J., Blei, D. M., and Zhu, X. (2007). A topic model for word sense disambiguation. In *Proceedings of Emperical Methods in Natural Language Processing.*

[Boyd-Graber and Resnik, 2010] Boyd-Graber, J. and Resnik, P. (2010). Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation. In *Proceedings of Emperical Methods in Natural Language Processing.*

[Boyd-Graber et al., 2012] Boyd-Graber, J., Satinoff, B., He, H., and III, H. D. (2012). Besting the quiz master: Crowdsourcing incremental classification games. In *Proceedings of Emperical Methods in Natural Language Processing.*

[Boyd-Graber et al., 2006] Boyd-Graber, J. L., Nikolova, S. S., Moffatt, K. A., Kin, K. C., Lee, J. Y., Mackey, L. W., Tremaine, M. M., and Klawe, M. M. (2006). Participatory design with proxies: Developing a desktop-PDA system to support people with aphasia. In *international conference on Human factors in computing systems.*

[Chang et al., 2009] Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Proceedings of Advances in Neural Information Processing Systems.*

[Eidelman et al., 2012] Eidelman, V., Boyd-Graber, J., and Resnik, P. (2012). Topic models for dynamic translation model adaptation. In *Proceedings of the Association for Computational Linguistics.*

[Hardisty et al., 2010] Hardisty, E., Boyd-Graber, J., and Resnik, P. (2010). Modeling perspective using adaptor grammars. In *Proceedings of Emperical Methods in Natural Language Processing.*

[Hu and Boyd-Graber, 2012] Hu, Y. and Boyd-Graber, J. (2012). Efficient tree-based topic modeling. In *Association for Computational Linguistics.*

[Hu et al., 2011] Hu, Y., Boyd-Graber, J., and Satinoff, B. (2011). Interactive topic modeling. In *Proceedings of the Association for Computational Linguistics.*

[Hu et al., 2012] Hu, Y., Zhai, K., Williamson, S., and Boyd-Graber, J. (2012). Modeling images using transformed Indian buffet processes. In *Proceedings of International Conference of Machine Learning*.

[Ma et al., 2009] Ma, X., Boyd-Graber, J., Nikolova, S. S., and Cook, P. (2009). Speaking through pictures: Images vs. icons. In *ACM Conference on Computers and Accessibility*.

[Nguyen et al., 2012] Nguyen, V.-A., Boyd-Graber, J., and Resnik, P. (2012). SITS: A hierarchical nonparametric model using speaker identity for topic segmentation in multiparty conversations. In *Proceedings of the Association for Computational Linguistics*.

[Sayeed et al., 2012] Sayeed, A. B., Boyd-Graber, J., Rusk, B., and Weinberg, A. (2012). Grammatical structures for word-level sentiment detection. In *North American Association of Computational Linguistics*.

[Zhai et al., 2012] Zhai, K., Boyd-Graber, J., Asadi, N., and Alkhouja, M. (2012). Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce. In *Proceedings of World Wide Web Conference*.

University of Maryland
E-mail address: jbg@umiacs.umd.edu