Research Statement

Hanna Wallach
http://www.cs.umass.edu/~wallach/
wallach@cs.umass.edu

1 Introduction

We are surrounded by complex social processes. From communications between friends, colleagues, or political leaders to the activities of corporate or governmental organizations, complex social processes underlie almost all human endeavor. As an active researcher and leader in the emerging interdisciplinary field of computational social science, my fundamental research goal is to develop new mathematical models and computational tools for understanding and reasoning about complex social processes using uncertain information. As a result, my research contributions are situated at the intersection of computer science, statistics, and the social sciences.

Computational social science is inherently interdisciplinary. To make truly groundbreaking advances, collaboration is necessary: social scientists are uniquely positioned to identify the most pertinent and vital research questions, as well as to provide insight into data sources and acquisition methods, while computer scientists and statisticians contribute significant expertise toward developing novel mathematical models and computational tools. As a result, most of my research projects are collaborations with researchers and practitioners in the social sciences, including political science, science of science policy, management science, and sociology. The interdisciplinary nature of my work is also reflected in my recent sources of funding, which include the National Institutes of Health (NIH) and the Office of Juvenile Justice and Delinquency Prevention (OJJDP), as well as the National Science Foundation (NSF) and the Intelligence Advanced Research Projects Activity (IARPA).

In order to establish productive collaborations between researchers with backgrounds in traditionally disparate research areas, it is imperative that researchers understand each other's cultural norms, research goals, and methodological frameworks. The first step in fostering this kind of understanding is establishing a common language for interdisciplinary communication. This task is well aligned with my wider goals as a professor. As described in my teaching statement, I strive to produce clear, unified presentations of complex, nuanced concepts. I have therefore devoted significant time to facilitating such communication by organizing two interdisciplinary workshops and a seminar series, as well as editing a journal special issue on computational social science. I have also published much of my recent work in a variety of emerging interdisciplinary venues in addition to established machine learning venues such as NIPS and ICML and high-profile scientific journals like *Nature Methods*. As a result of these activities, my first contribution as a professor was to develop a disciplined and organized way of characterizing and modeling complex social processes. This contribution is outlined in sections 1.1 and 1.2 below.

1.1 Characterizing Complex Social Processes

I maintain a broad definition of complex social processes: a complex social process consists of individuals or groups of individuals interacting with each other in order to achieve specific and often contradictory goals. Despite this wide definition, I believe that all complex social processes possess three commonalities: structure, content, and dynamics, as described below. Each of my research contributions focuses on one or more of these commonalities in a manner that transcends the specifics of that particular social process or application.

Structure: The structure of a complex social process is characterized by the interactions between individuals or groups of individuals. These interactions can be represented by a network, in which vertices represent individuals or groups and edges represent the interactions between them. Graph-theoretic properties—such as transitivity, reciprocity, and density—can then be used to understand and reason about the underlying social processes. For some social processes, interactions may not be directly observable and must be inferred from other information.

Content: The content associated with a complex social process is any information relating to or arising from the individuals or groups of individuals (vertices) and their interactions (edges). Depending on the social process, this information can take a variety of forms, including documents or other textual information, categories associated with individuals or the interactions between them, monetary information, citations, court proceedings, and voting records. Content can capture aspects of social processes that may not be apparent from structure alone.

Dynamics: Complex social processes exhibit temporal and spatial variation in both structure and content. For example, patenting practices often exhibit geographic variation, while scientific funding patterns and military interactions between countries are known to change over time. Studying these kinds of temporal and spatial dynamics, instead of focusing on single snapshots of structure and content, can reveal nuanced information that is crucial to understanding and reasoning about complex social processes, but might otherwise be neglected.

To date, much of the work in computational social science, particularly that originating in the social sciences and statistics, has focused on modeling interaction structure or, in some cases, structure and certain kinds of temporal dynamics. Studies that have considered content have mostly emphasized simple metadata or other low-dimensional information. My research has a different emphasis: leveraging textual content in order to facilitate nuanced analyses of complex social processes. Not only is textual content abundantly available and rich with nuanced detail, documents and other textual information are a key component in almost all social processes. Modeling content has many benefits, even when interaction structure is the primary object of interest. For many social processes, such as political influence networks, interaction structure may not be fully observed: there may be noisy or missing observations. In such cases, however, interaction structure may be indirectly evidenced via content. Content can also provide insight into the dynamics of complex social processes by revealing temporal or spatial patterns that may not be apparent from examining structure alone. Although much of my recent work has centered around modeling content and structure, my longer-term research directions all involve jointly modeling the structure, content, and dynamics of complex social processes in the face of uncertain or missing information.

1.2 Modeling Complex Social Processes

As a result of my research-related service activities, I have had the opportunity to shape an ongoing global discussion regarding commonalities and differences between the foci of computer scientists and social scientists. This discussion has led me to assert that the primary modeling goals pursued by researchers in computational social science are best divided into three categories: exploration, explanation, and prediction, as described below.

Exploration: Exploratory analyses uncover patterns in data—usually patterns that are not known to exist in advance. In other words, exploration is concerned with the question, "What do these data tell us that we don't already know?" Although exploration itself can be the ultimate modeling goal, exploration usually forms a precursor to explanatory or predictive analyses. Since most exploration tasks do not have an unambiguous right answer, evaluating the performance of models primarily intended for exploratory analyses can be non-trivial.

Explanation: Explanatory analyses find plausible or probable explanations for observations, i.e., answering "why" questions. The resultant explanations can then be compared with established social theories or other existing findings. As with models for exploration, the performance of models intended for explanatory analyses can be hard to evaluate unless the causes or explanations are known in advance, which is seldom the case.

Prediction: Finally, predictive analyses use observed data to make predictions about missing information or about future, yet-to-be-observed data. Even when prediction is not the ultimate modeling goal, predictive analyses can serve as a useful tool for validating models primarily intended for either exploratory or explanatory analyses.

Researchers with different backgrounds place varying emphases on these modeling goals. As described by King and Hopkins, "[C]omputer scientists may be interested in finding the needle in the haystack (such as a potential terrorist threat or the right web page to display from a search), but social scientists are more commonly interested in characterizing the haystack." [14] In other words, computer scientists are primarily interested in prediction, while social scientists are primarily interested in exploration and explanation. Articulating and examining such differences in modeling goals and cultural norms is imperative to the success of interdisciplinary collaborations. As an interdisciplinary researcher, it is seldom beneficial to pursue exploration, explanation, or prediction in isolation. My views are consistent with those of Schrodt, who argues that models for explanation must be validated via predictive analyses, while the utility of predictive models will be limited if they yield no insight into the characteristics that enable accurate predictions [26]. Exploratory analyses, though less well established in either computer science or the social sciences, yield crucial insights into observed data and its role in explanation or prediction. As a result, my research agenda involves projects with exploratory, explanatory, and predictive goals.

2 Methodological Framework

My primary methodological framework is that of machine learning—in particular, Bayesian latent variable modeling. Machine learning, one of the fastest-growing areas of computer science, is uniquely positioned to act as a framework for addressing the challenges involved in understanding and reasoning about complex social processes. In addition to possessing solid statistical foundations, machine learning is widely applicable to a diverse range of problems in social network analysis, time series analysis, information retrieval, and text analysis. Bayesian latent variable models use probabilities to represent beliefs and make consistent inferences given these beliefs, even under uncertain information. This use of probabilities requires that modeling assumptions, or prior beliefs, are made explicit, yielding mathematically rigorous and easily interpretable models. As a result, Bayesian latent variable models are not just "black-box" prediction tools: their interpretability makes them powerful and flexible models for exploratory and explanatory analyses as well as prediction. By developing such models, my research contributes to both machine learning and Bayesian statistics, as well as computational social science.

Regardless of the ultimate modeling goal, there are a number of mathematical and computational challenges that arise when analyzing the structure, content, and dynamics of complex social processes. Some of these challenges are theoretical, while others are more directly linked to real-world impact. My research agenda involves a blend of both theoretical and applied components: the needs of practitioners can drive theoretical work by encouraging researchers to explore the interplay between theory and practice, and to question previously unquestioned theoretical assumptions. My work is therefore heavily influenced by the relationship between the philosophical principles underlying Bayesian latent variable models and their applications. My statistics expertise means that I am skilled at developing tractable models and efficient inference algorithms for reasoning about missing or uncertain information. Via my collaborations with social scientists, I am able to incorporate pertinent domain knowledge into these models. Finally, my background in computer science means that I am extremely well placed to handle computational challenges such as aggregating and representing large quantities of unstructured data from sources with disparate emphases. Although this kind of mathematically rigorous, interdisciplinary work can be ambitious and time consuming, I believe it will ultimately yield high-impact, wide-ranging contributions.

3 Modeling Textual Content

My postdoctoral and early professorial research addressed a range of problems relating to the analysis of structured and unstructured textual information used in or arising from complex social processes. This work primarily focused on a class of Bayesian latent variable models known as statistical topic models [3]. These models automatically infer groups of semantically related words, known as topics, from word co-occurrence patterns within documents. Automated topic inference is extremely useful for characterizing the semantic content of document collections so large that manual human judgment is cost prohibitive. The resultant topics can be used to aid a variety of tasks such as detecting emergent areas of innovation, tracking topic trends across languages, and

identifying thematic collaborative communities. A selection of my research in these areas is described below.

3.1 Navigating and Discovering Funded Research

Information on research funding is important to various entities, including policy analysts, financial regulators, and advocacy organizations, as well as funding agencies and principal investigators. In particular, there is an urgent need for information on grants from NIH—the world's largest single source of biomedical research funding—because of its numerous awards (over 80,000 per year) and its complex organizational structure. NIH's twenty-five grant-awarding institutes and centers have distinct but overlapping missions, and the relationship between these missions and the research they fund is multifaceted. Because there is no comprehensive scheme that characterizes NIH research, understanding and navigating the NIH funding landscape can be a challenging task.

In collaboration with researchers and practitioners from NIH, ChalkLabs, University of California Irvine, and UMass Amherst, I was involved in the creation of NIHMaps: a publicly available exploration tool for NIH grants, intended to facilitate navigation and discovery of NIH-funded research.[1] NIHMaps uses statistical topic modeling to infer topics from grant titles and abstracts. The resultant topics are used to produce a two-dimensional visual output, in which grants are grouped according to their overall topic- and word-based similarity to one another. As a result, NIHMaps provides a framework that reflects scientific research rather than NIH administrative and categorical designations. We found that topic-based categories are not strictly associated with the missions of individual institutes, but instead cut across the NIH, albeit in proportions consistent with each institute's mission and policies. The graphical layout reveals a global research structure that is logically coherent but only loosely related to institute organization. Obtaining similar information without NIHMaps would require extensive exploration of institute websites, followed by time-consuming research on appropriate keywords for queries. NIHMaps offers an alternative approach that enables rapid and reproducible discovery of meaningful categorical information. This work was published in *Nature Methods* [28] and featured in a *Nature Methods* editorial [6].

3.2 Characterizing the Quality of Topic Models

Statistical topic models are extremely well suited to exploratory analyses. However, when using such models as a platform on which to build reliable analysis tools for practitioners and decision-makers, it must be possible to compare model features and innovations. Unlike supervised machine learning methods, most statistical topic models are not intended for prediction and do not rely on labeled data. Evaluating model performance is therefore non-trivial: the inferred topics cannot be compared to some set of known, "true" topics. For some downstream applications, such as information retrieval or document classification, there may be extrinsic prediction tasks on which performance can be evaluated; however, there is still a need for universal methods that measure model quality in ways that are accurate, computationally efficient, and independent of any particular application.

A natural evaluation metric for topic models is their predictive probability of previously unseen documents. This metric provides a clear and interpretable way of evaluating performance: the higher the probability, the better the model. Exact computation of this probability is intractable; however, several approximate estimation methods have been proposed in the topic modeling literature. With David Mimno, Iain Murray, and Ruslan Salakhutdinov, I experimentally analyzed a number of commonly used estimation methods and demonstrated that they are inaccurate and can change the relative ranking of different models [32]. We proposed two alternative methods that are accurate and efficient, thereby providing a new standard for evaluating and comparing statistical topic models.

Successful adoption of statistical topic models by social scientists and practitioners depends critically on their perceived reliability and topic quality. In practice, for real document collections, the inferred topics will be dominated by words that occur with high frequency but carry little information. Although this problem can be alleviated by removing such words prior to topic inference, discarding data in favor of designing models that "do the right thing" is clearly undesirable. This observation inspired me to formalize and examine some of the

[1] `http://nihmaps.org/`

fundamental assumptions, properties, and behaviors of statistical topic models. Most statistical topic models have relied on two prior beliefs: 1) In any set of documents, all topics are equally likely a priori, and 2) in any topic, all words are equally likely a priori. With Andrew McCallum and David Mimno, I investigated these previously unchallenged assumptions and explored alternative sets of prior beliefs [31]. Our work demonstrated that semantic coherence of inferred topics is significantly improved if some topics are more likely than others a priori, while there is no benefit to assuming that some words are more likely a priori. Most notably, we found that this combination of prior beliefs prevents topics from becoming dominated by high frequency, low content words. To take advantage of this discovery, David Mimno and I developed a novel algorithm for learning a priori topic probabilities [29]. This algorithm is over an order of magnitude faster than existing methods for learning such probabilities and adds negligible computational cost beyond standard inference algorithms for topic models.

While developing NIHMaps, as described in section 3.1, we repeatedly found that expert perceptions of the quality of inferred topics differed significantly from those of the researchers involved. Andrew McCallum, David Mimno, and I therefore worked with a program director at NIH to create a taxonomy of incoherent topics—i.e., topics that contain groups of words that are not semantically related—as well as a novel evaluation metric that automatically identifies such topics [23]. This coherence metric is extremely good at identifying topics that have been flagged by NIH experts as being poor quality, and was therefore instrumental in the subsequent development of NIHMaps.

3.3 Selecting Clustering Models for Social Science

Nonparametric Bayesian approaches to clustering obviate the need to specify a fixed number of clusters, thereby providing a robust mechanism for handling uncertainty about model structure. However, prior beliefs—such as assumptions about the propensity to create new clusters—still play a role in the clustering process. These beliefs are encoded as probability distributions, known as prior distributions. Since different prior distributions impose different properties on the resultant data partitions—such as the number of clusters or distribution of cluster sizes—their effects must be compared in order to select the most appropriate model for a given application [30].

In an interdisciplinary collaboration with political scientists Justin Grimmer and Frances Zlotnik, my student Rachel Shorey and I used a variety of prior distributions for nonparametric Bayesian clustering to define a new class of semiparametric statistical topic models for political texts. These models simultaneously infer the topics expressed in political texts as well as thematic partitions of those texts. We then used our models to develop a new, substance-driven methodology for social science model selection. This methodology combines the strengths of statistical approaches and substantive human judgment, thereby enabling social scientists to identify the models that are most appropriate for their explanatory needs. In order to incorporate expert judgment, we introduced several experimental methods that allow us to elicit careful evaluations of our models from domain experts. Using 19,000 press releases issued by members of the US House of Representatives in 2010, we used our models along with our selection methodology to show that ideologically extreme representatives dominate policy debates—a finding with widespread consequences for policy deliberation and lawmaking. This work was presented at the Society for Political Methodology's Summer Methods Meeting in 2011 [8] and at the Midwest Political Science Association Conference in 2012 [9]. We are now preparing to submit this work to a political science journal.

4 Modeling Structure

My recent research contributions have focused on jointly modeling the structure and textual content of complex social processes. These projects draw upon my previous work in statistical topic modeling, as well as ideas from Bayesian network modeling, to address a variety of exploratory, explanatory, and predictive goals. The impact of this work is wide ranging: these projects include collaborators in the intelligence community and the criminal justice system, as well as in machine learning and the social sciences. Two research directions are described below.

4.1 Creating Principled Visualizations

Much of the recent machine learning work on modeling the structure of complex social processes has focused on tasks such as predicting missing links or community membership. However, as described in section 1.2, social scientists are often more concerned with exploratory and explanatory analyses. A common approach to exploring and explaining the structure of complex social processes is to use some kind of statistical network model to perform quantitative analyses, while visualizing network structure using multidimensional scaling or a force-directed layout algorithm. Such visualization algorithms can produce extremely appealing pictures. However, if modeling and visualization are undertaken separately, the resultant visualizations may not directly reflect the model and its relationship to the observed data. Rather, these visualizations provide a view of the model and the data seen through the lens of the visualization algorithm and its associated assumptions, so any conclusions drawn from such visualizations can be biased by artifacts of the visualization algorithm. Although this viewpoint is beginning to gain traction within the visualization community, producing principled visualizations of the structure and content of social processes, i.e., visualizations that have precise interpretations in terms of an associated statistical model and its relationship to the observed data, remains an open challenge [7]. Prioritizing and tackling this open challenge is therefore one of my longer-term research objectives.

As a first step towards accomplishing this objective, Bruce Desmarais (political science) and I, along with my students Peter Krafft (now at MIT) and Juston Moore, recently developed a new Bayesian admixture model for discovering and visualizing topic-specific subnetworks in email data sets or other similarly structured communication networks. Rather than taking a two-stage approach in which subnetworks are discovered using one model and visualized using another, we present a single probabilistic model that partitions an observed email network into topic-specific subnetworks while simultaneously producing a visual representation of each subnetwork. As a result, our model is capable of producing principled visualizations of email networks, i.e., visualizations that have precise mathematical interpretations in terms of an underlying network model and its relationship to the observed data. This work has been presented at a variety of computational social science workshops—including the Workshop on Information in Networks [18], New Directions in Analyzing Text as Data [19], and the Annual Political Networks Conference [20]—and will appear at NIPS in late 2012 [17].

4.2 Nominating "Interesting" Vertices

In 2011, I was an invited participant in the ten-week-long annual SCALE (Summer Camp for Advanced Language Exploration) workshop, run by the Johns Hopkins University's Human Language Technology Center of Excellence (JHU HLTCOE).[2] This workshop brought together researchers from academia, industry, and government to work together on a variety of research problems relating to the task of vertex nomination, defined by Coppersmith and Priebe as follows: given a set of "interesting" individuals belonging to some complex social process, how can the structure and content of that social process be used to suggest and prioritize other similarly interesting individuals? [5] The Enron email corpus [16]—which comprises the inboxes and outboxes belonging to a set of individuals alleged to have engaged in fraudulent behavior—provides a real-world example of such a scenario. While many of these emails are innocuous, some document fraudulent activity. Given the set of emails (which are not labeled as fraudulent or otherwise) along with the partial network structure defined by the set of sender–recipient pairs, vertex nomination consists of constructing a ranked list of additional individuals who may have engaged in fraudulent behavior. Not only does this task have obvious relevance to both the intelligence and law enforcement communities, it is also related to a variety of advertising and social recommendation tasks.

My contributions to the SCALE workshop involved identifying and using relevant textual information in order to improve nomination performance based on network structure alone. Glen Coppersmith, Mark Dredze, and I drew upon ideas from supervised and unsupervised statistical topic modeling in order to develop and compare a variety of methods for identifying and exploring the textual content most likely to be associated with interesting individuals. Although this work is not yet published, the workshop was deemed to be success by its sponsors, and I am in the process of establishing a contract with the JHU HLTCOE in order to continue this collaboration.

[2] `http://hltcoe.jhu.edu/summer-camp-for-advanced-language-exploration-scale/`

Within law enforcement, one of the most important vertex nomination tasks is that of identifying and prioritizing contact offenders, i.e., individuals who directly and physically abuse children. Given the severity of these crimes, law enforcement cannot afford to identify contact offenders by waiting for victims or third parties to come forward; a proactive approach is imperative. Police investigations of child pornography trafficking within the US indicate that hundreds of thousands of individuals actively distribute child pornography on peer-to-peer file sharing networks. Furthermore, police logging of peer-to-peer activities does not require a warrant [15]. Investigating child pornography trafficking is critical to law enforcement: it is estimated that 16% of individuals arrested in cases that began with allegations or investigations of child pornography possession were found to be contact offenders [34]. However, with peer-to-peer logging information on hundreds of thousands of individuals, law enforcement needs ways of suggesting and prioritizing the subset of individuals worthy of further investigation.

I am currently involved in a three-year interdisciplinary collaboration to develop new methods for using peer-to-peer file-sharing information to nominate and prioritize peer-to-peer users that pose a high risk of being contact offenders. This project, which is funded by the OJJDP, involves computer scientists from UMass Amherst, sociologists from University of New Hampshire, as well as practitioners from the Internet Crimes Against Children Task Force and other law enforcement agencies. Together with our students, Brian Levine, Marc Liberatore, and I are working to combine logging information on peer-to-peer file-sharing (collected by police using software designed at UMass Amherst) with information about known contact offenders provided by sociologists at the University of New Hampshire in order to develop efficient predictive methods for suggesting potential contact offenders. These methods will ultimately be deployed by law enforcement agencies in order to assist with resource prioritization. Due to the illegal nature of child pornography, very little is known about the community of individuals who traffic such material. We are also investigating the feasibility of using file-sharing information along with respondent-driven sampling methods [13] to estimate characteristics of the hidden population involved in child pornography trafficking. We anticipate that these characteristics, which include the structure of the associated social network, will inform the development of new models for vertex prediction. Although our ultimate goal is prioritizing potential contact offenders, our focus is not "black-box" prediction. Rather, we hope to provide researchers and law enforcement agencies with a better understanding of the characteristics that tend to be indicative of contact offenders. As a result, it is imperative that our prediction models are interpretable in order to facilitate complementary exploratory and explanatory analyses.

5 Modeling Dynamics

My longer-term research directions are concerned with jointly modeling the dynamics, structure, and content of complex social processes. In particular, I am interested in models that go beyond simple parameterization of time series data and focus on 1) the dynamics of information transfer between entities and 2) the durations for which social processes remain in particular states, as well as the properties of those states and the conditions surrounding their transitions. As with my other projects, these projects build upon ideas from statistical topic modeling and Bayesian network modeling, however, they also draw inspiration from survival analysis, stochastic point processes, and Bayesian changepoint detection. A selection of these directions are described below.

5.1 Modeling Government Communication Networks

The changing structures of organizational communication networks are critical to collaborative problem solving [22]. While some questions may be answered by studying the structure of a communication network as a whole, other, more nuanced, questions can only be answered at finer levels of granularity—specifically, by studying topic-specific subnetworks. For example, breaks in communication (or duplicated communication) about particular topics may indicate a need for some form of organizational restructuring. Bruce Desmarais (political science) and I recently established a multi-year interdisciplinary project to build new Bayesian latent variable models for jointly analyzing intra-organizational government communication networks along with related external textual information. This project aims to study the topic-specific dynamics of government officials' communications, by focusing on the ways in which extra-governmental sources are related to 1)

intra-governmental communications and 2) publicly available government documents, including regulatory and legislative texts, as well as budgets, minutes and agendas. In this way, we will track the migration of specific topics to, within, and from government. Our models will build upon existing logistic–normal models of topic dynamics [2], as well as published and unpublished work on Gaussian Markov random field topic models by David Mimno and myself [24]. To facilitate exploratory and explanatory analyses of communication structure, we will also build upon my recent work on discovering and visualizing topic-specific subnetworks, described in section 4.1. As such, we will be able to characterize the structure and content of the democratic process at a fine-grained, topic-specific level. This characterization could help improve the efficiency with which local governments manage public health needs, address environmental risks, and establish revenue and spending policies. We are currently seeking funding from NSF to support this research for the next three years.

Although it is seldom possible for researchers to directly observe complete organizational communication networks, email provides one means by which such communication networks can be at least partially observed. As a result, email data sets hold the potential to answer many important scientific and practical questions within the organizational and social sciences. We have already undertaken extensive data acquisition work in order to obtain public record email archives from county governments in Florida and North Carolina, and are currently working to gather additional related documents, including regulatory and legislative texts, as well as minutes from county legislatures. As a pilot study, we recently introduced and analyzed a new public record email data set relevant to researchers in the organizational and social sciences as well as to machine learning researchers [17]. This data set consists of almost 2,000 emails between thirty managers of the departments that constitute the executive arm of government at the county level for New Hanover County, North Carolina. In this semi-autonomous local government, county managers act as executives, and the individual departments are synonymous with the individual departments and agencies in, for instance, the US federal government. Therefore, not only does this data set offer a view into the communication patterns of the managers of New Hanover County, but analyses of it also serve as case studies in modeling inter-agency communications in the US federal government administration.

5.2 Predicting Longevity using Textual Content

Many research questions in the social sciences involve either a direct or indirect interest in change and temporal dynamics. These longitudinal questions are often centered around the duration for which the structure or content of some complex social process persists in some state prior to a change of interest. As a result, survival analysis [4] is a common modeling framework within the social sciences. Statistical models for survival analysis capture the relationship between duration (or survival time) and potentially relevant underlying characteristics. Some survival analysis models facilitate prediction about missing or yet-to-be-observed durations, while others are intended primarily for exploratory or explanatory analyses. When applied to complex social processes, the characteristics of potential relevance are typically structural, or relate to simple metadata or other low-dimensional information—textual content is ignored. In machine learning, the majority of research involving temporal dynamics focuses on characterizing or making predictions about time series data, with applications including financial modeling, speech processing, and medical informatics—duration modeling remains largely unexplored.

One of my longer term research objectives is to develop new Bayesian latent variable models that capture the relationship between textual content and longevity. Since the textual content associated with a complex social process can reveal nuanced information that may not be apparent from its structure or from other content information, such models have the potential to transform survival analysis within the social sciences. Furthermore, the relative absence of survival analysis in machine learning means that such models have the potential to introduce the machine learning community to a whole new range of models and research questions.

With Bruce Desmarais and Rachel Shorey, I have undertaken pilot work on a new Bayesian modeling framework that combines ideas from survival analysis and statistical topic modeling in order to jointly model duration and textual content. This framework was initially developed in order to provide government secrecy researchers with new tools for studying the classification duration of formerly classified government documents. The US government protects a massive amount of secret data as part of its security classification system. This information is expensive to protect and maintain. In order to keep citizens informed as well as to keep costs down, the government is constantly releasing newly declassified documents to the public: according to the Information

Security Oversight Office, in 2011, human readers reviewed 52.8 million pages of information for declassification, of which 26.7 million were subsequently declassified, and $11.36 billion was spent on administration of the US government classification system [25]. Scholars interested in studying government history and policies on transparency and secrecy face a daunting task in examining even a small portion of these documents. Our modeling framework provides social scientists and classification policy decision-makers with new perspectives on the vast quantities of information belonging to the US government classification system. In 2011, we presented this work at the interdisciplinary Text as Data Conference [27]. We are currently working on a machine learning conference submission. In the longer term, I plan to extend this framework so as to facilitate the study of durations in other complex social processes, such as the length of time a patent is under review by the US Patent Office.

5.3 Detecting Changes in Structure and Textual Content

Although many social science research questions involve explicitly modeling duration, others are more concerned with recognizing when some complex social process has undergone either a transient or persistent change in state. There has been considerable work within the statistics and machine learning communities on detecting transient and persistent changes—respectively known as anomalies and changepoints—usually in the kinds of low-dimensional time series data found in medicine, finance, and even petroleum geology. In most of this work, the focus is on detecting either anomalies or changepoints, not both. Recently, there's been an increase in work (mostly originating in statistics and the social sciences) on detecting either anomalies or changepoints in complex social processes (e.g., [12]). This work has almost exclusively focused on changes in interaction frequency and structure. I am currently developing a new Bayesian modeling framework for detecting and characterizing changes in textual content, as well as changes in interaction structure. I am explicitly interested in changes characterized by an absence or decrease in certain interactions or topics, in addition to those characterized by an increase. Rather than detecting either anomalies or changepoints in isolation, my focus is on jointly detecting both.

Anomalies and changepoints are highly interrelated: when observed over some extended period of time, a complex social process is likely to undergo persistent changes in state. Any corresponding statistical model can capture these changes via appropriate changes to the model parameters. It is only with respect to the current state, i.e., model parameters, that observations are anomalous. If enough such observations accrue, the model parameters are arguably no longer capable of adequately explaining the data, and must be changed accordingly. In other words, a burst of anomalous observations often precedes a changepoint. I am developing new Bayesian latent variable models that capture this relationship between anomalies and changepoints. By combining ideas from anomaly and changepoint detection [12, 1, 33] and stochastic processes [10, 11] with ideas from statistical topic modeling, these new models will enable the detection and characterization of anomalies and changepoints in interaction structure and textual content. These models are of interest to researchers and practitioners in machine learning, statistics, and the social sciences, as well as the intelligence and law enforcement communities.

Drawing on well-established ideas from Kuhn [21], who characterizes paradigm shifts in science as changepoints preceded by anomalies, I am currently focusing my attention on detecting and characterizing scientific emergence. Policy-makers are increasingly interested in understanding the social processes that shape the emergent phases of scientific discoveries and technological inventions. While there has been extensive research on the later stages of scientific development, much less is known about the factors surrounding emergence. This information could help scientific policy-makers stimulate breakthrough research. I am interested in analyzing changes in scientific publication and patenting behavior, reflected in their textual content, and modeling the ways in which these changes are driven by "social" factors, such as collaborations, organizational involvement, funding sources, and geographical locations. To this end, I am an investigator on a five-year award from IARPA, in collaboration with researchers from Raytheon BBN Technologies, SciTech Strategies, UMass Amherst, and the University of Pennsylvania. This award is part of IARPA's Foresight and Understanding from Scientific Exposition Program.

6 Summary

My fundamental research goal is to develop new mathematical models and computational tools for understanding and reasoning about the content, structure, and dynamics of complex social processes. My work is inherently interdisciplinary: truly groundbreaking advances in computational social science will only be made if researchers and practitioners in computer science, statistics, and the social sciences strive to establish productive collaborations. Over the past two years, I have successfully pursued and established such interdisciplinary collaborations. As a result, I am now poised to make important research contributions at the intersection of these fields. This approach to computational social science has the potential to transform our understanding of the complex social processes that underlie society, and I am proud to be expanding the frontiers of this nascent area.

References

[1] R. Adams and D. MacKay. Bayesian online changepoint detection. arXiv:0710.3742v1, 2007.

[2] D. Blei and J. Lafferty. Dynamic topic models. In *Proceedings of the Twenty-Third International Conference on Machine Learning*, 2006.

[3] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] J. Box-Steffensmeier and B. Jones. *Event History Modeling: A Guide for Social Scientists*. Cambridge University Press, 2004.

[5] G. Coppersmith and C. Priebe. Vertex nomination via content and context. arXiv:1201.4118v1, 2012.

[6] Editorial. Mapping the money. *Nature Methods*, 7(6):437, 2011.

[7] S. Fienberg. A brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics*, 22, 2012.

[8] J. Grimmer, R. Shorey, **H. Wallach**, and F. Zlotnik. A class of Bayesian nonparametric topic models for measuring expressed priorities in political texts. In *Proceedings of the Society for Political Methodology Twenty-Eigth Annual Summer Meeting*, 2011.

[9] J. Grimmer, R. Shorey, **H. Wallach**, and F. Zlotnik. A class of Bayesian semiparametric topic models for political texts. In *Proceedings of the Midwest Political Science Association Conference*, 2012.

[10] A. Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B*, 33:438–443, 1971.

[11] A. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58:83–90, 1971.

[12] N. Heard, D. Weston, K. Platanioti, and D. Hand. Bayesian anomaly detection methods for social networks. *Annals of Applied Statistics*, 2010.

[13] D. Heckathorn. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 1997.

[14] D. Hopkins and G. King. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247, 2010.

[15] O. Kerr. *Computer Crime Law*. West, second edition, 2009.

[16] B. Klimt and Y. Yang. Introducing the Enron corpus. In *Proceedings of the First Conference on Email and Anti-Spam*, 2004.

[17] P. Krafft, J. Moore, B. Desmarais, and **H. Wallach**. Topic-partitioned multinetwork embeddings. In *Advances in Neural Information Processing Systems Twenty-Five*, 2012.

[18] P. Krafft, J. Moore, **H. Wallach**, , B. Desmarais, and J. ben Aaron. Topic-specific communication patterns from email data. In *Workshop on Information in Networks*, 2012.

[19] P. Krafft, J. Moore, **H. Wallach**, , B. Desmarais, and J. ben Aaron. Topic-specific communication patterns from email data. In *Proceedings of the Third New Directions in Analyzing Text as Data Conference*, 2012.

[20] P. Krafft, J. Moore, **H. Wallach**, B. Desmarais, and J. ben Aaron. Modeling government email networks. In *Proceedings of the Fifth Annual Political Networks Conference*, 2012.

[21] T. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, third edition, 1996.

[22] W. Mason and D. Watts. Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, 109(3):764–769, 2012.

[23] D. Mimno, **H. Wallach**, M. Leenders, E. Talley, and A. McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.

[24] D. Mimno, **H. Wallach**, and A. McCallum. Gibbs sampling for logistic normal topic models with graph-based priors. In *Proceedings of the 2008 Neural Information Processing Systems Workshop on "Analyzing Graphs"*, 2008.

[25] I. S. O. Office. Annual report to the president, 2011.

[26] P. Schrodt. Seven deadly sins of contemporary quantitative political analysis. In *Proceedings of the Annual American Political Science Association Meeting and Exhibition*, 2010.

[27] R. Shorey, **H. Wallach**, and B. Desmarais. Toward a framework for the large-scale textual and contextual analysis of government information declassification patterns. In *Proceedings of the Second Annual Text as Data Conference*, 2011.

[28] E. Talley, D. Newman, D. Mimno, B. H. II, **H. Wallach**, G. Burns, M. Leenders, and A. McCallum. Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods*, 8(6):443–444, 2011.

[29] **H. Wallach**. *Structured Topic Models for Language*. PhD thesis, University of Cambridge, 2008.

[30] **H. Wallach**, S. Jensen, L. Dicker, and K. Heller. An alternative prior process for nonparametric Bayesian clustering. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.

[31] **H. Wallach**, D. Mimno, and A. McCallum. Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems Twenty-Two*, 2009.

[32] **H. Wallach**, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proceedings of the Twenty-Sixth International Conference on Machine Learning*, 2009.

[33] R. Turner, Y. Saatci, and C. Rasmussen. Adaptive sequential Bayesian change point detection. In *Proceedings of the 2000 Neural Information Processing Systems Workshop on "Temporal Segmentation"*, 2009.

[34] J. Wolak, D. Finkelhor, and K. Mitchell. Child-pornography possessors arrested in Internet-related crimes: Findings from the NJOV study. Technical report, National Center for Missing and Exploited Children, 2005.