

Machine Learning for Computational Social Science

University of Massachusetts Amherst, Department of Computer Science

Hanna M. Wallach (Director)

<http://www.cs.umass.edu/~wallach/>

wallach@cs.umass.edu

Our primary research goal is to develop new mathematical models and computational tools for analyzing vast quantities of structured and unstructured data in order to identify and answer scientific questions about complex social processes. The social processes that we study are diverse and include those that underlie free and open source software development, scientific collaboration, and the US political system. To this end, we work on techniques for aggregating and representing large amounts of information from data sources with disparate emphases, methods for analyzing relational and social network data, efficient algorithms for inference, and robust methods for reasoning under uncertainty. Machine learning, one of the fastest-growing areas of computer science, is uniquely positioned to act as a framework for developing such methods and tools because of its solid theoretical foundations in statistics and widespread applicability to diverse problems in social network and text analysis. Our research contributes to machine learning, Bayesian statistics, and, in collaboration with social scientists, to the nascent field of computational social science. Our projects include:

Government Information Declassification Patterns: The US government protects a massive amount of secret data as part of its Security Classification System. In order to keep citizens informed, as well as to keep costs down, the government is constantly releasing newly declassified documents to the public: human readers manually declassified almost 29 million pages of information in 2009 alone. Scholars interested in learning about the history and policy of government transparency face a daunting task in examining even a small portion of these documents. In order to facilitate the process of learning about these documents, we are investigating the content of government-released, formerly-classified documents created throughout the twentieth century. Our investigation has two dimensions: 1) We use survival analysis to consider questions relating to time, such as when documents were created and how long they tend to remain classified, and 2) we model document content using statistical topic models. We are developing new statistical models that draw upon mathematical ideas from both survival analysis and statistical topic modeling to jointly model temporal and content information, so that each can inform inferences about the other. These techniques have the potential to contribute to a greater understanding for social scientists and classification policy-makers.

Cross-Language Data Collections: Evaluating and tracking the rate and direction of scientific and technological progress in other countries is a critical step towards determining national scientific priorities and assessing related policy decisions. However, this kind of cross-national, thematic analysis is computationally challenging: even if policy-makers have access to scientific documents from other countries, the content of these documents will be represented very differently if these documents are written in different languages. As a result, scientific policy-makers need automated tools that characterize the semantic content of documents in many languages. To date, statistical topic models have been primarily used in monolingual, or at most bilingual, contexts. In collaboration with others at UMass, we developed a polylingual topic model that automatically infers topics (groups of semantically-related words) aligned across multiple languages. Unlike previous bilingual topic models, our model does not rely on expensive, sentence-aligned translations, but requires only relatively small numbers of semantically-comparable texts (i.e., documents that are not translations of one another, but are very likely to be about similar subject matter, such as Wikipedia articles or software documentation). This flexibility enables social scientists and policy-makers to perform data-driven analysis of semantic similarities and differences across multiple languages in a wide variety of settings, and can even be used to analyze semantically-comparable documents in a single language, such as pairs of patents and publications that describe the same scientific or technological innovation.

Understanding Diversity of Science: The study of science and innovation policy has traditionally focused on the ways in which particular policy actions impact the rate of scientific progress. Although productivity is an important way of characterizing inventive activity, a solid understanding of the relationship between policy and scientific progress requires a deeper investigation of the ways in which policy decisions shape the diversity of science. We focus on two types of diversity: idea diversity—the array of different ideas that arise from discoveries and inventions—and individual diversity—the variety of people and organizations contributing to scientific progress. To explore the relationship between policy actions and diversity, we are developing new methods and tools that will better enable social scientists and policy-makers to quantify scientific diversity, as well as to characterize and assess the impact of policy actions on diversity, and balance issues of

effectiveness, productivity, and social participation in the context of scientific growth. This line of work involves aggregating and representing information from multiple data sources with different emphases (e.g., publications, patents, and grants) and developing predictive models for large-scale collaboration networks.

Free and Open Source Software Development Communities: Free and open source software (FOSS) is a rapidly-growing movement, with communities around the world working together to develop software that may be freely copied, distributed, studied, and improved. Over the past decade, there has been considerable commercial, noncommercial, and academic interest in FOSS, due to its unique position in bringing together technological, legal, and social structures to provide a foundation for collaboration on a massive scale. Despite this interest, the complex organizational and social processes surrounding FOSS remain largely unknown. These processes form a compelling area of collaborative study for social and computer scientists. Due to the open and distributed nature of FOSS development, most interaction between developers takes place online, via mailing lists, Internet Relay Chat (IRC), version control systems, and software collaboration tools. To date, few studies of FOSS development have made use of this immense quantity of publicly available data. We are mining and analyzing these diverse data sources in order to better study the organizational and social processes that underlie FOSS development. This task is extremely challenging: unlike documents written for more formal purposes, the text arising from FOSS development communities is highly unstructured. Consequently, significant text analysis is required prior to developing models that can answer social science questions. For example, IRC channels typically have multiple interleaved conversations occurring at any point in time. Conversational thread disentanglement—i.e., determining which utterances belong to which conversation—must be performed prior to analyzing of IRC logs. Data-driven study of FOSS development communities is a rich source of challenging research problems for both computer and social scientists, and our interests in computational social science, combined with our machine learning expertise and involvement in FOSS development, make this a particularly compelling long-term research direction.

Lab Members

Hanna Wallach (Director): In fall 2010, Hanna Wallach started as an assistant professor in the Department of Computer Science at the University of Massachusetts Amherst. She is one of five core faculty members involved in UMass's computational social science research initiative. Prior to this, Hanna was a postdoctoral researcher, also at UMass, where she developed statistical machine learning techniques for analyzing complex data regarding communication and collaboration within scientific and technological innovation communities. Hanna's Ph.D. work, undertaken at the University of Cambridge, introduced new methods for statistically modeling text using structured topic models—models that automatically infer semantic information from unstructured text and information about document structure. In addition to her many papers on statistical machine learning techniques for analyzing structured and unstructured data, Hanna's tutorial on conditional random fields is extremely widely cited and used in machine learning courses around the world. Her recent work (with Ryan Adams and Zoubin Ghahramani) on infinite belief networks won the best paper award at AISTATS 2010. As well as her research, Hanna works to promote and support women's involvement in computing. In 2006, she co-founded an annual workshop for women in machine learning, in order to give female faculty, research scientists, postdoctoral researchers, and graduate students an opportunity to meet, exchange research ideas, and build mentoring and networking relationships. In her not-so-spare time, Hanna is a member of Pioneer Valley Roller Derby, where she is better known as Logistic Aggression.

Rachel Shorey is in the third year of her Ph.D. She is particularly interested in using machine learning methods to understand government policies and politics, with the ultimate goal of producing easy-to-understand information to help increase popular participation in government. When she is not in front of her computer, Rachel can be found volunteering as a Girl Scout leader or knitting sweaters.

Anton Bakalov is a second-year M.S./Ph.D. student. He is interested in developing probabilistic graphical models for analyzing the thematic structure of document collections. Currently, he is focusing on models that capture topic hierarchies. In his spare time he likes to play tennis, billiards and basketball.

Meagan Day is a first-year M.S. student. Her research interests include improving the robustness, usability, and interpretability of statistical models for text analysis and of the software used to implement such models. Meagan's extracurricular activities include playing upright bass in a band, cooking, and aquarium-keeping.

Peter Krafft completed his undergraduate degree in mathematics and statistics at UMass Amherst before moving into the computer science department's M.S. program. He currently works with Prof. Sridhar Mahadevan in the field of representation discovery and with Prof. Hanna Wallach in nonparametric Bayesian statistics. Though he doesn't know what he does in his spare time, according to his girlfriend, he likes drinking fine beer, sleeping on his comfortable bed, and petting the adorable kitties that live with him.