Abstract: Print and Probability: A Statistical Approach to Clandestine Publication

Print and Probability is a National Science Foundation-funded project that develops novel machine learning and computer vision techniques to infer thousands of book and pamphlet printers whose identities have eluded scholars for several hundred years. Before the modern era, the book trade was often dangerous and secretive. For fear of persecution and punishment, printers between 1473-1700 declined to attach their names to about a quarter of all known books and pamphlets. In the period between 1660-1688, the proportion of anonymous and clandestine publications sometimes reaches as high as 35\%.
However, now that roughly 115,000 books have been digitized by the Early English Books Online (EEBO) project, defects and variations in the printing tools of this era may hold the key to identifying these printers. Once an individual piece of metal type is damaged, it creates unique stamps. Since typesets belonged to specific printers, impressions of damaged type can thus serve as the fingerprints to identify the printers of tens of thousands of clandestine publications. The Print and Probability project automatically detects and tracks these unique pieces of damaged type in order to uncover new information about the history of books. The methods developed in this project could be generalizable to other important tasks and domains - for example, digital forensics and authorship attribution. In addition, the Print and Probability project trains students in a multidisciplinary way, engaging them in collaboration across multiple fields.

By developing new techniques for visual anomaly detection, the Print and Probability project detects damaged letterforms that create consistent aberrations. Based on these damaged type extractions, the project develops probabilistic models of both printer and damaged letter form identifications that allow direct inference of printers at scale. This framework also incorporates other sources of evidence into the identification model - most significantly, the spelling, punctuation, and spacing habits of individual press-house compositors, whose distinctive practices lend themselves to clustering and automatic attributions across all pages of text in the collection. Integrating a new method for automatic compositor attribution, this project develops a statistical model for printer identification that leverages the same sources of evidence compiled manually by scholars of rare books, but at a scale and speed never before possible.