# An Ontology of Privacy Law Derived from Published Articles Using Latent Dirichlet Allocation Probabilistic Topic Modeling

**Robert Sprague**
Department of Management & Marketing
University of Wyoming
Laramie, WY 82071
*spraguer@uwyo.edu*

**Nicole Barberis**
IBM
Laramie, WY
*nbarberis@yahoo.com*

## Abstract

This paper presents preliminary results from applying latent Dirichlet allocation probabilistic topic modeling algorithms to a document collection comprised of published law review articles from 1891 through 1970, and sample publications from 1980, 1990, 2000, and 2010—all citing to Warren's and Brandeis's *The Right to Privacy* and substantively discussing privacy law issues. Our initial interpretation of the results reveals a few important trends: early discussion of issues associated with the tort of appropriation of name or likeness; fourth amendment government searches and surveillance; and emerging trends in information privacy associated with data collection, as well as the rise of the Internet. Once our initial data collection has been completed, by filling out the years 1971 through 2012, we hope to not only confirm these initial observations but also reveal additional privacy law trends, developing an ontology of privacy law based upon the topics revealed in published law review and journal articles.

## 1 Introduction

Privacy, being an evolutionary product of social development[3], has been a human need and desire for millennia. Privacy law scholarship, in contrast, is a relatively recent phenomenon. Within this recent profusion of scholarship lies a conundrum: there is no clear definition of privacy[5]; there is not even consensus of what would constitute an adequate description. Fundamental concepts associated with privacy have been identified and analyzed—for example, seclusion, intimacy, surveillance, anonymity, and control of information. But, as Helen Nissenbaum has noted, most calls for privacy arise from context, as well as advancing technologies[4], meaning the legal system often has difficulty identifying and protecting rights to privacy. Without a coherent construction of privacy principles shared by the community of scholars, the legal discipline will never explicitly articulate those principles[5].

This paper reports preliminary results from a research project aimed at identifying fundamental privacy law principles derived from the writings of legal scholars and commentators using probabilistic topic modeling. A latent Dirichlet allocation (LDA) process, which identifies sets of terms that more tightly co-occur, is incorporated into the topic modeling analysis to identify words most closely associated with each identified topic. The LDA therefore provides insight into the context in which each identified topic occurs.

Most published law review articles that cite to Samuel Warren's and Louis Brandeis's

seminal article, *The Right to Privacy* [6] (some 3000 articles), are being converted to plain text. *The Right to Privacy* was selected as the focal point of the document collection because it is the original published scholarly call for a formal legal right to privacy in the United States; hence, the vast majority of privacy law publications cites to it. Probabilistic topic modeling using latent Dirichlet allocation is being applied to the document collection in time slices to reveal the evolution of fundamental privacy law concepts expressed in the legal literature published from 1890 through 2012. The ultimate goal of this project is to identify the fundamental conceptual structure of privacy law in the United States as reflected by over a century of published law review and journal articles.

# 2 Methodology

The initial document collection will be comprised of relevant published law review and journal articles that cite to Warren's and Brandeis's *The Right to Privacy*, as identified within the Westlaw and HeinOnline collections. All documents selected for the collection are converted to plain text. In addition, all titles, author names, section headings, footnotes/endnotes, and supplemental materials are removed in an effort to create a collection limited to addressing substantive privacy law issues. At present, only approximately 20% of the anticipated initial collection, representing privacy law articles published up to and through 1970, has been converted. The current document collection has been divided into the following time-slice corpora: 1891-1940 (which includes *The Right to Privacy*), 1941-1950, 1951-1960, and 1961-1970. A cumulative corpus has also been created for the time period 1891-1970. For this paper, additional partial corpora were created from articles published in 1980, 1990, 2000, and 2010. Work is continuing to build complete corpora for: 1971-1980, 1981-1990, 1991-2000, 2001-2010, and 2010-2012.

The LDA topic model algorithms are then applied to the corpora using MALLET[2]. The critical MALLET output files used in this project include the following files: keys, weights, words count, and composition.

## 2.1 Results

Identifying significant privacy law topics from the individual corpora can be approached from different views of the MALLET results. The MALLET output files were parsed and analyzed using R, which also generated visualizations of the data.

### 2.1.1 Ubiquity Measure

Our "Ubiquity Measure" is an attempt to visualize the frequency of occurrence of the term "privacy" throughout the individual corpora. Within each time-slice corpus, every occurrence of the term "privacy" was assigned an "AB" weight (where A = the normalized weight of the topic in which the term occurs and B = the normalized weight of the term within its topic). The AB weight is therefore the adjusted weight of the term depending on the weight of the topic within which it occurs. Our "Ubiquity Measure" is the sum of all the AB values within a particular time-slice corpus. In effect, the "Ubiquity Measure" represents the degree of occurrence of the term privacy in each time-slice relative to all the other time slices.

### 2.1.2 Privacy Constellations

Our "Privacy Constellations" reflect the normalized weight of the term privacy within each topic in which it occurs within each time-slice corpus. This not only reflects the other terms co-occurring with the term privacy within a topic, but also each term's relative weight within the topic. This allows one to see the context in which the term privacy was used by published authors within each time-slice corpus.

### 2.1.3 Treemaps

We created Treemaps for each time-slice corpus reflecting the weight of the topics in the corpus and labeled each area with the first term in the topic's cluster because the "discovered" topic is mostly about that first word. This is a good way to see which ideas were important in an era, reflected by their relative weight.

### 2.1.4 Topics and Terms

The topics and associated terms identified for each time slice are reflected in a 20x20 matrix built from MALLET's "keys" output file. However, these matrices can be somewhat cumbersome to read and interpret. We believe our Ubiquity Measure, Privacy Constellations, and Treemaps offer more helpful visualizations of our data for interpreting the results.

## 2.2 Initial Interpretations of the Preliminary Results

Due to the page limitation for this proposal—targeting the Application theme of the NIPS Topic Models workshop—visualizations of our initial results are not included. If this proposal is accepted into the workshop, the visualizations can comprise a significant portion of this project's presentation.

We can make a few generalizations from the preliminary results. Looking at the Treemaps, it is not surprising to find the term "privacy" as one of the (normalized) heaviest-weighted "top" terms. The Treemaps also provide a few additional insights: for example, the term property was a heavily weighted term through 1950, but then subsequently drops out of the "top" terms. Meanwhile, the terms government, fourth, and amendment make sizable appearances in the 1961-1970 corpus, implying much greater attention to fourth amendment privacy rights related to government searches.[1]

In the 1891-1940 corpus, the term "privacy" is closely associated with the terms "public," "publication," "picture," "person," "interest," "invasion," "news," "life," "advertising," and "photograph." This clustering reflects a focus in the early development of privacy law on the ability of individuals to control their images and likenesses—fundamentally, the tort of appropriation of name or likeness. As the document collection is completed beyond 1970 it will be interesting to observe how prevalent this topic will remain in comparison to other "privacy" topics.

The term "information" appears quite frequently within the corpus, although its most frequent appearance is in a very low-weighted topic and most frequently co-occurs with the terms "data," "credit," "computer," "personal," "privacy," "bureau," "access," "system," and "files" (we can infer "credit" and "bureau" refer to credit bureaus because those two terms most tightly co-occur also in in the same topic). These terms can be associated with the growing computerized collection of personal information that began in the 1960s. The fact that many of these terms, particularly "information," appear frequently in the corpus, but most frequently in a very low-weighted topic, indicates that this privacy issue was a late blooming topic, at least for this corpus, but when it did appear, it was discussed quite extensively.

The term "amendment" is one of the most frequently occurring terms, yet it occurs most frequently in relatively low-weighted topics. And when "amendment" does appear in those topics, it is associated with the terms "fourth," "electronic," "eavesdropping," "privacy," "surveillance," "evidence," "justice," "telephone," "agent," "conversation," "search," "seizure," "warrant," and "arrest." While these term imply discussions of fourth amendment privacy rights vis-à-vis government searches and surveillance, the terms "griswold" and "connecticut" also appear with "amendment," indicating discussion of *Griswold v. Connecticut*[1], in which the Supreme Court held that a Connecticut law forbidding the use of contraceptives unconstitutionally intruded upon the right of marital privacy.

Based on the 1891-1970 document corpus, three major "areas" of privacy can be discerned in the published literature: rights associated with one's name or likeness, fourth (and more generally ninth and fourteenth) amendment rights against government searches and surveillance, and the emerging issue of information privacy in a rapidly computerizing society.

# 3 Conclusion

[1] Treemaps were not created for the partial single-year corpora (1980, 1990, 2000, and 2010) due to their limited expanse over time.

This paper has presented preliminary results from applying LDA probabilistic topic modeling algorithms to a document collection comprised of published law review articles from 1891 through 1970, and sample publications from 1980, 1990, 2000, and 2010—all citing to Warren's and Brandeis's *The Right to Privacy* and substantively discussing privacy law issues. Our initial interpretation of the results reveals a few important trends: early discussion of issues associated with the tort of appropriation of name or likeness; fourth amendment government searches and surveillance; and emerging trends in information privacy associated with data collection, as well as the rise of the Internet. Once our initial data collection has been completed, by filling out the years 1971 through 2012, we hope to not only confirm these initial observations but also reveal additional privacy law trends, developing an ontology of privacy law based upon the topics revealed in published law review and journal articles.

## Acknowledgments

The authors are extremely grateful for the support and assistance they have received from: Professor Ken Gerow, Chair, University of Wyoming Department of Statistics.

## References

[1] Griswold v. Connecticut, 381 U.S. 479 (1965).

[2] McCallum, A. (2002) *MALLET: A Machine Learning for Language Toolkit* (2002), http://mallet.cs.umass.edu.

[3] Moore, B., Jr. (1984) *Privacy: Studies in Social and Cultural History*. New York, NY: Pantheon Books.

[4] Nissenbaum H. (2010) *Privacy in Context*. Stanford, CA: Stanford Law Books.

[5] Roberts, L.A. (1993) *The Ontology of Privacy*. (unpublished Ph.D. dissertation, University of Oregon) (on file with authors).

[6] Warren, S.D. & Brandeis, L.D. (1890) The Right to Privacy. *Harvard Law Review* **4**(5): 193-220.