One Hundred Twenty Years of U.S. Privacy Law Scholarship: A Latent Semantic Analysis

Robert Sprague, *University of Wyoming*
Kevin Grauberger, *University of Wyoming*
Nicole Barberis
ONE HUNDRED TWENTY YEARS OF U.S. PRIVACY LAW SCHOLARSHIP: A LATENT SEMANTIC ANALYSIS

by

Robert Sprague, Kevin Grauberger, and Nicole Barberis

Privacy, like an elephant, is perhaps more readily recognized than described.1

I. INTRODUCTION

Privacy, being an evolutionary product of social development,2 has been a human need and desire for millennia. Privacy law scholarship, in contrast, is a relatively recent phenomenon. Of all the privacy-related law review articles published in the history of the United States, for example, over ninety percent were published after 1990—and half that amount in the past decade. Within this recent profusion of scholarship lies a conundrum: there is no clear definition of privacy;3 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

---

1 John B. Young, Introduction: A Look at Privacy, in PRIVACY 1, 2 (John B. Young ed., 1978).
3 See Richard B. Parker, A Definition of Privacy, 27 RUTGERS L. REV. 275, 275-76 (1974) (“Currently, there is no consensus in the legal and philosophical literature on a definition of privacy.”); Daniel J. Solove, A Taxonomy of Privacy, 154 U. PA. L. REV. 477, 485-86 (2006) (“Privacy is too complicated a concept to be boiled down to a single essence. Attempts to find such an essence often end up being too broad and vague, with little usefulness in addressing concrete issues.”); Leilani A. Roberts, The Ontology of Privacy 7 (Oct. 1993) (unpublished Ph.D. dissertation, University of Oregon) (on file with author) (“[T]here is substantial agreement among scholars that ‘privacy,’ as a bare concept, may not permit a clean, clear definition.”).
4 See generally HELEN NISSENBAUM, PRIVACY IN CONTEXT (2010) (arguing that a framework of contextual integrity can determine when people will perceive new information technologies and systems as threats to privacy); Ken Gormley, One Hundred Years of Privacy, 1992 WIS. L. REV. 1335 (1992) (arguing that “seeking a simple definition of privacy is a misguided quest,” as “historical jolts” or “catalysts” produce “new brands of privacy when existing law is incapable of dealing with unexpected societal and technological changes”).
privacy. Without a coherent construction of privacy principles shared by the community of scholars, the legal discipline will never explicitly articulate those principles.

This paper reports 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, which is comprised of a suite of algorithms that attempt to discover hidden thematic structures in large archives of documents. Topic modeling algorithms are statistical methods that analyze the words of texts to discover topics (themes) contained within, how those topics are connected to each other, and how they change over time. A latent Dirichlet allocation 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 latent Dirichlet allocation 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 (over 2500 articles) have been converted to plain text to build a corpus of publications focused on the law of privacy. 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 has been applied to the document collection in time slices to reveal the evolution of fundamental U.S. privacy law concepts expressed in the legal literature published from 1890 through 2013. Studies in different disciplines have demonstrated the ability of latent Dirichlet allocation to analyze the rich underlying structures of a particular domain—depicting emerging and sustained trends in a given discourse. 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.

---

5 See, e.g., Hon. William H. Rehnquist, Is an Expanded Right of Privacy Consistent with Fair and Effective Law Enforcement? Or: Privacy, You’ve Come a Long Way, Baby, 23 U. KAN. L. REV. 1, 2 (1974) (arguing that “in most situations in which claims to privacy are urged, there are two sides to the issue, and, if the balance is struck in favor of ‘privacy,’ some other societal value will suffer”); Lior Jacob Strahilevitz, Toward a Positive Theory of Privacy Law, 126 HARV. L. REV. 2010, 2010 (2013) (arguing that “[p]rivacy protections create winners and losers[,]” and “that understanding the identities of privacy law’s real winners and losers is indispensable both to clarifying existing debates in the scholarship and to helping predict which interests will prevail in the institutions that formulate privacy rules”).

6 See Roberts, supra note 3, at 9.


8 See Blei, supra note 7, at 77-78.

9 See id.

This paper first presents an overview of the origins of privacy law to provide a backdrop for the motivations underlying this project. It then provides an overview of the topic modeling process using latent Dirichlet allocation to explain the underlying analytical methodology. Results from applying the statistical modeling to a collection of published law review and journal articles are then presented and discussed—specifically, fundamental topics identified from published articles in the following time slices: 1891 – 1950, 1951 – 1960, 1961 – 1970, 1971 – 1980, 1981 – 1990, 1991 – 2000, 2001 – 2013, and the full corpus, 1891 – 2013. Our analysis reveals that privacy law in the United States comports most closely with the Georgia Supreme Court’s 1905 description of privacy from the seminal case *Pavesich v. New England Life Insurance Company*: “the right of a person to be secure from invasion by the [government or] public into matters of a private nature.”

II. ORIGINS OF PRIVACY LAW

What is privacy? And, in particular, what legal rights to privacy does U.S. law protect? What legal rights to privacy should U.S. law protect? This conundrum has vexed legal scholars and commentators for over a century. There have been attempts to describe, more than define, privacy. Warren and Brandeis, of course, most famously asserted the right to privacy as the right to be let alone. In 1905, the Georgia Supreme Court described the right of privacy as “the right of a person to be secure from invasion by the public into matters of a private nature.” In 1960, after examining some 300 court cases, William Prosser described, rather than defined, privacy as not one tort, but a complex of four:

The law of privacy comprises four distinct kinds of invasion of four different interests of the plaintiff, which are tied together by the common name, but otherwise have almost nothing in common except that each represents an interference with the right of the plaintiff. . . “to be let alone.” Without any attempt to exact definition, these four torts may be described as follows:

1. Intrusion upon the plaintiff’s seclusion or solitude, or into his private affairs.
2. Public disclosure of embarrassing private facts about the plaintiff.
3. Publicity which places the plaintiff in a false light in the public eye.
4. Appropriation, for the defendant’s advantage, of the plaintiff’s name or likeness.

---

12 50 S.E. 68, 71-72 (Ga. 1905); see infra Part III.C.

13 Warren & Brandeis, supra note 10, at 195. Warren and Brandeis were quoting THOMAS M. COOLEY, TREATISE ON THE LAW OF TORTS OR THE WRONGS WHICH ARISE INDEPENDENT OF CONTRACT 29 (1880) (stating that “[t]he right to one’s person may be said to be a right of complete immunity: to be let alone”). “No right is held more sacred, or is more carefully guarded, by the common law, than the right of every individual to the possession and control of his own person, free from all restraint or interference of others, unless by clear and unquestionable authority of law.” Union Pac. Ry. Co. v. Botsford, 141 U.S. 250, 251 (1891) (affirming the lower court’s denial of Union Pacific’s motion to surgically examine Botsford, who had brought a personal injury action against Union Pacific).

14 *Pavesich*, 50 S.E. at 71-72.

15 William L. Prosser, *Privacy*, 48 CAL. L. REV. 383, 388-89 (1960) (citation omitted). These four categories were subsequently compiled in the RESTATEMENT (SECOND) OF TORTS (1976) at §§ 652B (Intrusion upon Seclusion),
A few years later, Alan Westin described privacy as “the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others.” Westin identified four states of privacy: (1) solitude, where an individual is separated from the group and freed from the observation of others; (2) intimacy, where an individual exercises seclusion to achieve a closer relationship with others; (3) anonymity, the freedom from identification and surveillance in public places or while performing public acts; and (4) reserve, the creation of a psychological barrier against unwanted intrusions. More recently, Daniel Solove developed a taxonomy of privacy—focusing on activities that create problems—identifying four groups of harmful activities: (1) information collection, (2) information processing, (3) information dissemination, and (4) invasion. Solove later identified six types of privacy: (1) the right to be let alone; (2) limited access to the self; (3) secrecy; (4) control over personal information; (5) personhood; and (6) intimacy.

In the formative years of the United States, eavesdropping was the principal form of privacy concern—preventing someone from standing under the eaves of a home and listening to the conversation inside. However, by the late nineteenth century, two interrelated developments threatened individuals’ privacy—the first was “keyhole journalism,” reflecting the growth of newspapers printing salacious material; the second was the advent of technologies such as audio recording devices and instantaneous photography. These developments piqued one individual in particular:

---

652C (Appropriation of Name or Likeness), 652D (Publicity Given to Private Life), and 652E (Publicity Placing Person in False Light).

16 ALAN F. WESTIN, PRIVACY AND FREEDOM 7 (1967).

17 Id. at 31-32.


20 See DAVID J. SEIPP, THE RIGHT TO PRIVACY IN AMERICAN HISTORY 2 (1978) (discussing how home construction made listening to private conversations possible). Eavesdropping was considered an actionable public nuisance in England by the beginning of the nineteenth century. See SIR WILLIAM BLACKSTONE, COMMENTARIES ON THE LAWS OF ENGLAND 168-69 (Thomas B. Wait, & Co. 1807) (classifying eavesdropping as one of seven wrongs whose punishment is short of death). “Eaves-droppers, or such as listen under walls or windows or the eaves of a house, to hearken after discourse, and thereupon to frame slanderous and mischievous tales, are a common nuisance and presentable at... [court].” Id. Eavesdropping was recognized as an actionable wrong by American courts by the early nineteenth century. See, e.g., Commonwealth v. Lovett, 4 Pa. L.J. Rpts. (Clark) 226, 226 (Pa. 1831) (“No man has a right to pry into your secrecy in your own house.”); see also Lance E. Rothenberg, RE-THINKING PRIVACY: Peeping Toms, Video Voyeurs, and Failure of the Criminal Law to Recognize a Reasonable Expectation of Privacy in the Public Space, 49 AM. U. L. REV. 1127, 1140-42 (2000) (describing the origin of the term “peeping tom” as well as its early prosecutions as a breach of the peace).

21 FRANK LUTHER MOTT, AMERICAN JOURNALISM: A HISTORY OF NEWSPAPERS IN THE UNITED STATES THROUGH 250 YEARS 1690 TO 1940, at 444 (1941) (describing keyhole journalism as an “invasion of privacy by prying reporters”).

22 See ROBERT ELLIS SMITH, BEN FRANKLIN’S WEB SITE: PRIVACY AND CURIOSITY FROM PLYMOUTH ROCK TO THE INTERNET 123 (2000) (listing technological developments in the latter half of the nineteenth century); see also Robert E. Mensel, “Kodakers Lying in Wait”: Amateur Photography and the Right to Privacy in New York, 1885-
On January 25, 1883, [Samuel] Warren had married Miss Mabel Bayard, daughter of Senator Thomas Francis Bayard, Sr. They set up housekeeping in Boston’s exclusive Back Bay section and began to entertain elaborately. The *Saturday Evening Gazette*, which specialized in “blue blood items,” naturally reported their activities in lurid detail. This annoyed Warren, who took the matter up with [Louis] Brandeis.\(^{23}\)

The rest, as they say, is history.\(^{24}\) One can see reflected in *The Right to Privacy* concerns of intrusive technology recording and publishing intimate private life.\(^{25}\) Warren and Brandeis were searching for some form of legal recourse to prevent such intrusions. They recognized that traditional libel and slander torts would not address the type of harm they had in mind—“our law recognizes no principle upon which compensation can be granted for mere injury to the feelings.”\(^{26}\) Recognizing that “[t]he common law secures to each individual the right of determining, ordinarily, to what extent his thoughts, sentiments, and emotions shall be communicated to others,”\(^{27}\) Warren and Brandeis argued that traditional intellectual property

---

1915, 43 AM. Q. 24, 25 (1991) (discussing the impact of unprecedented technological change upon the public’s psyche).

\(^{23}\) ALPHEUS THOMAS MASON, BRANDEIS: A FREE MAN’S LIFE 70 (1946). Warren’s father-in-law, Thomas Francis Bayard, was a U.S. Senator, one-time presidential candidate, and Secretary of State under Grover Cleveland. LEWIS J. PAPER, BRANDEIS 35 (1983). There is some dispute as to what specific event triggered Warren’s call to action. Prosser reported it was “when the newspapers had a field day on the occasion of the wedding of a daughter, and Mr. Warren became annoyed.” Prosser, *supra* note 15, at 383 (citing MASON, *supra*, at 70 (stating that the *Saturday Evening Gazette* reported the Warrens’ activities in “lurid detail”)). But see James H. Barron, *Warren and Brandeis, The Right to Privacy, 4 Harv. L. Rev. 193 (1890): Demystifying a Landmark Citation*, 13 SUFFOLK U. L. REV. 875, 894 (1979) (suggesting the triggering event was the reporting of the wedding of a cousin since the Warrens’ oldest daughter could not have been more than seven years old in 1890; asserting the public disclosure that there were “no bridesmaids” was the embarrassing incident that “launched a thousand lawsuits” (paraphrasing Prosser, *supra* at 423 (suggesting that Warren’s daughter “was the face that launched a thousand lawsuits”)). A Brandeis biographer suggests Warren called upon Brandeis to write the article with him after becoming “outraged when photographers invaded his babies’ privacy and snapped perambulator pictures.” ALFRED LIEF, BRANDEIS: THE PERSONAL HISTORY OF AN AMERICAN IDEAL 51 (1936). See also Amy Gajda, *What if Samuel D. Warren Hadn’t Married a Senator’s Daughter?: Uncovering the Press Coverage that Led to the Right to Privacy*, 2008 Mich. St. L. Rev. 35 (2008) (concluding, after examining some sixty newspaper stories reporting on the personal lives of Warren and his family, that *The Right to Privacy* would not have been written if Warren had not married into a political family).

\(^{24}\) See Barron, *supra* note 23, at 876-77 (noting the article has “assumed a hallowed place in both legal literature and history” and the “near unanimity among courts and commentators that the Warren-Brandeis conceptualization created the structural and jurisprudential foundation of the tort of invasion of privacy”).

\(^{25}\) “Instantaneous photographs and newspaper enterprise have invaded the sacred precincts of private and domestic life; and numerous mechanical devices threaten to make good the prediction that ‘what is whispered in the closet shall be proclaimed from the house-tops.’” Warren & Brandeis, *supra* note 10, at 195; see also Gajda, *supra* note 23 (“Warren had the misfortune . . . of marrying into a highly newsworthy political family and for years thereafter suffered the occasional burdens of public curiosity. Had he not, it is quite likely that ‘The Right to Privacy’ would not have been written.”).


\(^{27}\) *Id.* at 198.
rights, such as copyright, were insufficient.\textsuperscript{28} Instead, they argued, “the right to privacy, [w]as a part of the more general right to the immunity of the person,—the right to one’s personality.”\textsuperscript{29}

It is not as if the notion of privacy did not exist in the United States prior to the publication of The Right to Privacy. As discussed above, eavesdropping had long been recognized as an actionable harm.\textsuperscript{30} In 1811, a Louisiana court prohibited a newspaper editor from publishing the contents of a letter without permission from the sender,\textsuperscript{31} and in an 1881 case in which a doctor allowed an unmarried man with no medical training to be present while a woman gave birth, the court acknowledged the woman’s right to privacy during “a most sacred” occasion.\textsuperscript{32}

But while Warren and Brandeis are recognized for putting into motion a formal judicial recognition of the right to privacy,\textsuperscript{33} they did not end the discussion of exactly what constitutes a legally protected right to privacy. As Diane Zimmerman argues, “Lawyers and philosophers have generated a vast literature on the subject without being able to agree upon some core of values or interests common to each of the cases in which the ‘right to privacy’ has been applied.”\textsuperscript{34} Zimmerman concludes, “A part of the difficulty is that the phrase today is a catch-all, attached to a broad range of interests which often have little or nothing to do with the tort originally envisioned by Warren and Brandeis.”\textsuperscript{35} For example, Warren and Brandeis were concerned with intimate details being revealed to the public. What of today’s near-ubiquitous electronic tracking and surveillance? “Increasingly, what we are worried about are practices that involve collecting, using and disclosing information that is not sensitive or intimate and that is increasingly

\textsuperscript{28} Id. at 200 (“But where the value of the production is found not in the right to take the profits arising from publication, but in the peace of mind or the relief afforded by the ability to prevent any publication at all, it is difficult to regard the right as one of property, in the common acceptation of that term.”)

\textsuperscript{29} Id. at 207; see Neil M. Richards & Daniel J. Solove, Privacy’s Other Path: Recovering the Law of Confidentiality, 96 GEO. L.J. 123, 128-31 (2007) (succinctly analyzing Warren’s and Brandeis’s arguments in The Right to Privacy); Diane L. Zimmerman, Requiem for a Heavyweight: A Farewell to Warren and Brandeis’s Privacy Tort, 68 CORNELL L. REV. 291, 295 (1983) (asserting that Warren and Brandeis appeared “to believe that the details of one’s personal life ‘belonged’ in some sense to the individual and could not be ‘used’ by others without permission”).

\textsuperscript{30} See supra note 20 and accompanying text.

\textsuperscript{31} Dennis v. LeClerc, 1 Mart. (o.s.) 297 (Orleans 1811); see also Woolsey v. Judd, 11 How. Pr. 49 (N.Y. Super. Ct. 1855 (holding same); Richards & Solove, supra note 29, at 144 n.43 (describing nineteenth-century cases preventing the disclosure of confidential letters).

\textsuperscript{32} De May v. Roberts, 9 N.W. 146, 149 (Mich. 1881).

\textsuperscript{33} See Prosser, supra note 15, at 386 (impliedly crediting Warren and Brandeis for the fact that by 1960 the right to privacy was recognized in one form or another by an “overwhelming majority of the American courts”). However, Warren and Brandeis were not the first to publicly call for the recognition of a right to privacy. See, e.g., E. L. Godkin, The Rights of the Citizen. I.—To His Own Reputation, SCRIBNER’S MAG., July-Dec. 1890, at 58, 65 (claiming it was a “natural right” to decide how much the public may know of an individual’s personal thought and feeling, tastes, habits, and private doings and affairs, including those of the family living under one’s roof).

\textsuperscript{34} Zimmerman, supra note 29, at 294 & n.9 (citing various scholars struggling with the ideal of a definition of privacy).

\textsuperscript{35} Id. at 294-95. See also generally Gormley, supra note 4, at 1340 (arguing that Warren’s and Brandeis’s tort-based right to privacy is only one of five “distinct species” of privacy; the other four being Fourth Amendment privacy, First Amendment privacy, fundamental-decision privacy, and state constitutional privacy).
collected in public—concerns that do not easily fall within the domain of traditional privacy theory.\textsuperscript{36}

The frustration is clearly reflected in the literature. G. Robert Blakey wonders if privacy is “an idea of which there are two hundred definitions,”\textsuperscript{37} “better suited to literary than legal analysis.”\textsuperscript{38} One commentator opined nearly fifty years ago:

No other area in the law has caused so much confusion to the courts than that dealing with the right of privacy. The doctrine of privacy is still very much in its infancy seventy-five years after its conception. Only the barest outlines of its nature and extent have as yet been sketched by judicial decisions.\textsuperscript{39}

More recently, regarding being named with Daniel Solove to lead a project to compile a Restatement (Third) Information Privacy Principles, privacy scholar Paul Schwartz stated “that the project is designed to bring clarity to a body of common law that has become a bewildering array of conflicting and overwhelming rules differing from state-to-state and the federal government.”\textsuperscript{40}

There have been approximately 3500 law review articles published in the United States that cite to The Right to Privacy,\textsuperscript{41} and nearly 5000 law review articles contain the word “privacy” in their titles.\textsuperscript{42} Owing to the quantity—not to ignore quality—of publications on the topic, it is clear there is much to be said about privacy. It is not a settled, nor static, issue.

This paper lays the foundation, and presents data, not for a definition or categorization of privacy, but for an ontology of privacy—an interpretation of the scope of privacy without assuming that anything exists beyond written expressions; namely, from published law review and journal articles.\textsuperscript{43}


\textsuperscript{38} Id. (quoting William M. Beany, The Constitutional Right to Privacy in the Supreme Court, 1962 SUP. CT. REV. 212, 214 (1962); see also Young, supra note 1, at 4 (“Privacy means different things to different men. It is an elusive concept, difficult to define.”).)

\textsuperscript{39} Note, Right of Privacy—Property Rights—Unfair Competition—Constitutional Guarantees—Personal Rights, 11 N.Y. L.F. 120, 120 (1965) (footnote omitted). Indeed, D.N. MacCormick even argues there is a significant distinction between defining “privacy” and the “right to privacy.” D.N. MacCormick, Comment, Privacy: A Problem of Definition?, 1 BRT. J.L. & SOC’Y 75, 75 (1974).

\textsuperscript{40} David McAuley, ALI to Commence Restatement Project Compiling Information Privacy Principles, BLOOMBERG BNA ELECTRONIC COMM. & L. REP., 18 ECLR 684 (Apr. 10, 2013).


\textsuperscript{42} This is an admittedly crude method to calculate the number of published law review articles substantively discussing privacy, particularly, as indicated infra note 51, some articles may address privacy only tangentially.

\textsuperscript{43} Here we are borrowing liberally from Stanislaw Leśniewski’s system of ontology as a general principal of being. See John T. Kearns, The Contribution of Leśniewski, 8 NOTRE DAME J. FORMAL LOGIC 61, 62 (1967). In this context, “extensionality requires that any expression must be considered only in terms of those existing objects to
III. BUILDING AN ONTOLOGY OF PRIVACY USING PROBABILISTIC TOPIC MODELING

Probabilistic topic modeling, a refinement of latent semantic analysis, represents a computational method to understand digitized data. The basic intuition behind latent Dirichlet allocation (LDA), the simplest form of topic modeling, is that documents exhibit multiple topics. Latent Dirichlet allocation is a statistical model of document collections (corpora) that tries to capture this intuition. More formally, LDA is a generative probabilistic model of a corpus in which documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Stated another way, topic modeling provides “a probability distribution over words that picks out a coherent cluster of correlated terms.” See Appendix A for a conceptual and algorithmic description of the computational process.

For our project, this process is applied to document collections comprised of published law review and journal articles in order to discern fundamental privacy law concepts.

A. Methodology

The document collection is 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. The Right to Privacy was selected as the initial means to identify relevant articles because, as noted previously, it is recognized as the seminal scholarly article advocating the recognition of a formal legal right to privacy within U.S. law; which it can be related.” Id. at 72. Fundamentally, Leśniewski’s ontology addressed syntactic paradoxes based on a theory of distributive predication. See John Thomas Canty, Ontology: Leśniewski’s Logical Language, 5 FOUND. LANGUAGE 455, 455 (1969).

44 See Thomas K. Landauer et al., An Introduction to Latent Semantic Analysis, 25 DISCOURSE PROCESSES 259, 259 (1998) (“Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text. The underlying idea is that the aggregate of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other.”) (citation omitted).

45 See Blei, supra note 7, at 77.

46 Id. at 78.

47 Id.


49 Steyvers & Griffiths, supra note 7, at 425.

50 Warren & Brandeis, supra note 10.

51 Not every published article that cites to The Right to Privacy substantively addresses privacy; for example, some articles cite to The Right to Privacy as an example of a law review article’s influence on the development of law, while others may address privacy only as a minor, secondary topic. These articles have been excluded from the document collection. See infra Part III.D. (discussing limitations to the project with respect to culling articles). The articles selected for this project include scholarly articles, student notes and comments, and analyses of recent decisions.

52 See supra notes 33 & 41.
consequently a substantial number of published articles which address U.S. privacy law cite to *The Right to Privacy*. A total of 2539 documents selected for the collection were converted to plain text. In addition, all titles, author names, section headings, footnotes and endnotes, and supplemental materials were removed in an effort to create a collection limited to addressing substantive privacy law issues. The document collection has been divided into the following time-slice corpora: 1891 – 1950 (which includes *The Right to Privacy*), 1951 – 1960, 1961 – 1970, 1971 – 1980, 1981 – 1990, 1991 – 2000, 2001 – 2013, and a cumulative corpus for the time period 1891 – 2013. 

The LDA topic model algorithms were then applied to the corpora using the MALLET software implementation. Data, in the form of plain text, are first imported into MALLET. Once the data are imported, MALLET provides a number of output options. The critical output files used in this project include:

- A “keys” file that provides an overview of the topics identified within the data; it contains a “key” consisting of the top k words for each topic (where k is the number of most probable words for each topic after model estimation). This file also reports the Dirichlet parameter of each topic that will be roughly proportional to the overall portion of the collection assigned to a given topic—in other words, the weight of the latent topic and the words that compose that topic (the cluster of words that define the latent topic).

---

53 Citing to *The Right to Privacy* does not automatically qualify a published article for inclusion in the document collection. As noted supra, note 51, articles were reviewed for relevancy in an attempt to collect only documents that substantively address privacy law. It is also acknowledged that many privacy law articles will not necessarily cite to *The Right to Privacy*. Limiting document collection to articles citing to *The Right to Privacy* is a methodology used to identify a large collection of relevant documents. See infra Part III.D. 

54 This is a fairly daunting task, as the majority of law review articles published prior to 1990 are available only through HeinOnline, and stored as pdf replicas of the original articles. Conversion to plain text often leaves garbled words, as well as footnotes which must be removed on a page-by-page basis. In addition, all formatting codes—particularly paragraph codes—must be stripped from each document.

55 Fundamentally, the approach is to track the evolution of privacy law decade by decade; however, due to the relatively few articles published prior to 1950, the first time slice covers the sixty-year span from 1891 – 1950.

56 Comprising 159 documents.

57 Comprising 109 documents.

58 Comprising 160 documents.

59 Comprising 326 documents.

60 Comprising 427 documents.

61 Comprising 516 documents.

62 Comprising 842 documents. It was decided the time period 2011-2013 was too short to identify fundamental privacy concepts unique to that time period.

63 Comprising, as noted above, 2539 documents.

A “composition” file that lists the source documents and reports the rank of topics identified with each document.

The reported weight of each topic represents the strength of likelihood that the collection of words represented within the topic conveys a meaningful concept. A lower weight generally means that the words associated with the topic do not appear very frequently within the corpus or they co-occur less frequently. This notion is also reflected in the “composition” file where one generally sees more documents linked to higher weighted topics; in other words, a topic is more significant because it is discussed more frequently (using the words that appear in the topic), which supports a key concept behind this project—the essence of privacy is identified by the words of the authors themselves.

There are three variables that were used to define the parameters of each MALLET topic model “run.” The number of iterations was increased from the default of 1000 to 2000; this variable controls how many samples will be taken of the corpus. According to MALLET, increasing this variable can increase the quality of the topic model. We also chose to create topics that listed just the first twenty identified words. The MALLET “keys” file lists the words in each topic in descending order of frequency of occurrence; we found that for the most part the fifteenth to twentieth words reflected very low frequencies.

Perhaps the most critical variable to select in any topic modeling exercise is the number of topics to identify. There is no consensus as to the optimum number of topics. We experimented with different numbers of topics: 20, 25, 50, 100, 200, 250, and 400 topics. We determined that twenty-five topics appeared to be optimal; less than twenty-five topics underfit the data (i.e., we risked missing some topics), while models with fifty or more topics generally overfit the data—we found the topics were diffused, as if MALLET were trying to identify too many topics.

Following is a representation of a typical topic (from the 1891 – 1950 corpus):

1.13189 | privacy | picture | person | publication | invasion | consent | damages | photograph | action
| purposes | property | violation | advertising | recognized | personal | Brandeis | pictures | publicity | libel | newspaper.

The number is the weight assigned to this topic, followed by the identified terms in descending order of frequency of occurrence.

This is validated using another MALLET output file that records the number of times a specific word is associated (if at all) with each topic.

Techniques have been applied to confirm an optimum number of topics. For example, Massey et al. performed multiple topic models on a sample of their corpus and then calculated the perplexity—a measure of the predictive likelihood of a model of text—of the models. See Massey et al., supra note 11; see also Steyvers & Griffiths, supra note 7, at 441 (discussing various methods to determine the number of topics). See generally Airoldi et al., Reconceptualizing the Classification of PNAS Articles, 107 PROC. NAT’L ACAD. SCI. 20899 (2010) (explaining their methodology to determine twenty to forty topics were optimal for their data set, which comprised documents from multiple disciplines).

See Steyvers & Griffiths, supra note 7, at 441 (noting that “a solution with too many topics will result in uninterpretable topics that pick out idiosyncratic word combinations”). For the 2001 – 2013 corpus, we found that a 50-topic run did identify a few additional relevant topics (see infra notes 250-252 and accompanying text), though a 50-topic run of the 1891 – 2013 corpus did not identify any topics that were not already identified in the previous time-slice corpora.

Using the 2001 – 2013 corpus for an idiosyncratic example, at 50 topics a very weak topic (with the lowest weight of all the topics) appears:
We ran three topic models on each corpus. Since topic models are unsupervised, meaning no prior assumptions are made about the data, each model begins with a random assignment of topics that is then honed through each iteration. As a result, while they may be quite similar, no two topic models run against the same data will be identical, both in terms of the topic weights as well as the exact words comprising each topic. In order to identify what appeared to be the most relevant topics among the runs, we developed a scoring system that first identified words in each topic that appeared in all three runs. We then scored each topic, among the three runs, based on the number of three-word-matches, the weight of each topic, and how close the individual weights were to each other. The scoring was not the definitive determination of whether a topic was identified for our results; it did, however, significantly aid in identifying the results among the topics within the three runs. The topics reported below reflect the words that appeared only in all three runs.

Finally, we employed a stop words list (sometimes referred to as a stoplist) to filter out common prepositions and conjunctions. We customized the stop words list with a few additional generic legal terms such as plaintiff, defendant, law, case, court, held, etc.

---

| geisha | perp | jurors | walk | Iwasaki | gender | golden | housing | juror | tenant | women | transgender | sex | Wharton | tenants | trial | walks | dire | fiction | media |


geisha | Iwasaki | golden | narrative | common | force | house | hamlet | Japan | life | defense | book | fiction | author | coke | prostitution | act | sex | household | Japanese


---


61 The final score “number” is arbitrary as it is meant merely to rank the topics according to how strongly they fit the parameters.

62 See Murphy Choy, Effective Listings of Function Stop Words for Twitter, 3 Int’l J. Advanced Computer Sci. & Applications 8 (2012), available at http://arxiv.org/pdf/1205.6396 (“A stop words list refers to a set of terms or words that have no inherent useful information. Stop words create problems in identification of key concepts and words from textual sources when they are not removed due to their overwhelming presence both in terms of
B. Analysis of Results

As noted previously, LDA topic modeling does not identify single-word topics; rather, it identifies probable topics consisting of multiple, clustered terms. As such, identifying and reporting in this paper substantive privacy law topics from the corpora involve filtering with subject-matter expertise. Many highly weighted topics were often generic from a legal sense—for example, comprised of terms such as decision, supreme, rights, constitutional, etc., but not terms that could be directly linked to a privacy-related concept.73

Our identified topics are presented below within their respective time slices. We list and discuss the topics in descending order by weight,74 as well as list representative articles from which those topics were derived.75 As such, one can view the topics as ranked in order of prevalence within each slice, and the terms comprising each topic as ranked in order of frequency of occurrence with their respective topic.76

72 We did not, however, stem any words. As such, we see in our models singular and plural versions of some words, though one or the other has been removed from the results reported here. We are not yet at a point where we are confident we can stem certain words without possibly limiting the relevance of topics.

73 For example, the top-weighted topic derived from the 1891-1950 corpus is:

| statute | question | interest | action | basis | general | decision | found | purpose | rule | ground |
| protected | present | similar | judicial | involved | persons | holding | protection | result |

In other words, these terms frequently occurred in close proximity to each other within the 159 articles; however, they do not tell us anything with respect to privacy.

74 Specifically, we rank by “Run 1” weight. We ran three sequential runs for each corpus, numbering them runs 1, 2, and 3. Although, as noted above (see infra, text surrounding note 69), each run is unique in terms of topic weights and terms, each identified topic within each run is comparable to its sister topics in the other runs with respect to weight and terms.

75 We identify the relevant articles using the “Run 1” composition file, ranking the articles in descending order by the predominant topic with which they are associated. Each article may contribute terms to multiple topics; however, those contributions are weighted, so it is possible to list the articles within each topic that contributed most heavily to the terms within the topic. Our representative list of articles is not exhaustive for each topic; we limit to the first few articles that appear representative of the topic, often eliminating what appear to be duplicative and tangential discussions. The point of the representative articles is reflected in that they are “representative” of the discussion(s) giving rise to the topic. Unless otherwise noted, articles are listed according to The Bluebook: A Uniform System of Citation (19th ed.) Rule 1.4(i).

76 Again, the topics we report here reflect only the words contained in all three runs, and we have removed most singular/plural duplications.
1891 – 1950

The most heavily weighted topic identified from this corpus reveals that during the first sixty years, privacy law articles centered on the publication of pictures and photographs and particularly the unauthorized use of a person’s name or likeness; also included are concepts of general discussion vis-à-vis the developing right to privacy in the United States. In Roberson v Rochester Folding Box Company the New York Court of Appeals refused to recognize a common law right of privacy arising from the unauthorized publication of the plaintiff’s likeness on flour boxes. In contrast, the Georgia Supreme Court held that the unauthorized publication of person’s picture for commercial purposes constituted an invasion of privacy in Pavesich v. New England Life Insurance Company. As a result of Roberson, the New York legislature passed the country’s first privacy legislation, but it was limited to the commercial appropriation of a person’s likeness. As such, much of the early published commentary focused on the development of privacy case law vis-à-vis The Right to Privacy. For example, a Comment in a 1920 Cornell Law Quarterly contrasted Roberson with two Louisiana cases that recognized a right of privacy involving the publications of individuals’ pictures in a “rogues” gallery after they had been arrested but before conviction. The Comment summarized the state of the right

77 privacy | picture | person | publication | invasion | action | violation | recognized

78 See, e.g., in chronological order, Recent Decisions, Equity—The Right of Privacy, 1 COLUM. L. REV. 491 (1901) (discussing Roberson v. Rochester Folding Box Co., 171 N.Y. 538 (1902)); Recent Cases, Right of Privacy—Infringement—Unauthorized Use of Portrait for Advertising Purposes, 18 HARV. L. REV. 625 (1905) [hereinafter Unauthorized Use of Portrait for Advertising Purposes] (discussing Pavesich v. New England Life Ins. Co., 50 S.E. 68, 77-81 (Ga. 1905)); Note, Possible Interests in One’s Name or Picture, 28 HARV. L. REV. 689, 689 (1915) (“[I]t is not surprising that with the advent of modern photography and the growth of a certain type of unscrupulous journalism the law has come to recognize to a limited extent an individual’s right of privacy, a right not to have his personal affairs subjected to public comment and scrutiny without his consent. Yet the courts have handled the matter very unsatisfactorily.”) (footnotes omitted); Comment, Torts—Invasion of the Right of Privacy, 1 S. CAL. L. REV. 293 (1928); Recent Cases, Torts—Right of Privacy—Parents’ Right Not to Have Picture of Deceased Deformed Child Made Public, 79 U. PA. L. REV. 511 (1931); Recent Decisions, Torts—Libel—Photographs—Right of Privacy, 32 Mich. L. REV. 1172 (1934); Note, Torts—The Right of Privacy, 25 MINN. L. REV. 619 (1941); Note, Torts—The Right of Privacy in Kentucky, 38 Ky. L.J. 487 (1950).

79 Prosser notes “[i]n nearly every jurisdiction the first decisions were understandably preoccupied with the question whether the right of privacy existed at all, and gave little or no consideration to what it would amount to if it did.” Prosser, supra note 15, at 388; see also Current Decision, Is There a Right of Privacy in Colorado?, 9 ROCKY MTN. L. REV. 378 (1937) (“Squarely confronted with the newly developed legal doctrine of the right of privacy, the Supreme Court of Colorado was extremely timorous in discussing and recognizing the doctrine in Colorado in McCreey v. Miller’s Grocereta and evaded the issue entirely by deciding the case upon other grounds.”); involving a woman who’s photograph was used without her permission by a grocery chain to advertise coffee) (footnote omitted).

80 171 N.Y. 538 (1902).

81 50 S.E. 68, 77-81 (Ga. 1905); see also Edison v. Edison Polyform Mfg. Co., 67 A. 392 (N.J. Ch. 1907) (granting Thomas Edison an injunction against a company using his name in association with its business and products).

82 See N.Y. CIV. RIGHTS LAW § 50 (McKinney 2009) (“A person, firm or corporation that uses for advertising purposes, or for the purposes of trade, the name, portrait or picture of any living person without having first obtained the written consent of such person . . . is guilty of a misdemeanor.”) (held constitutional, Rhodes v. Sperry & Hutchinson Co., 85 N.E. 1097 (N.Y. 1908), aff’d, Sperry & Hutchinson Co. v. Rhodes, 220 U.S. 502 (1911)).

83 Warren G. Willsey, Comment, Equity: Right of Privacy: Injunction to Protect Personal Rights, 5 CORNELL L.Q. 177, 178 (1920).
to privacy in America as of 1920: “Without exception... all jurisdictions have refused to protect this so-called right of privacy, unless there has been also an infringement of a property right or a breach of trust and confidence upon which the court could base the right to injunctive relief.”

A second, lower weighted topic is derived almost exclusively from discussions and analysis of Sidis v. F-R Publishing Corporation, involving a child mathematics prodigy who had graduated from Harvard at age sixteen. Thirty years later, The New Yorker magazine published a true but unflattering story about the individual (Sidis), highlighting his modest work and living conditions. The Second Circuit Court of Appeals concluded it was “not yet disposed to afford to all of the intimate details of private life an absolute immunity from the prying of the press. Everyone will agree that at some point the public interest in obtaining information becomes dominant over the individual’s desire for privacy.”

One commentator summarized the state of privacy protection at this point for public figures as follows:

Though numerous cases involving the so-called right of privacy have arisen, the law remains in much confusion. The basic problem underlying the future development of the law of privacy is that of the definition of the right in view of the freedom of the press to reveal those matters with which the public has a legitimate concern.

A third identified topic is a continuation of the commercial use of a person’s name or likeness without authorization—with articles citing frequently to Atkinson v. Doherty, Pavesich, and Edison v. Edison Polyform Manufacturing Company. Closely related, but still a

---

84 Id.; see also Comment, Property Rights in Letters, 46 YALE L.J. 493 (1937).
85 public | interest | privacy | life | news | publication | private | individual | article | legitimate | figure | matters | privilege | recovery | commercial | concern
86 113 F.2d 806 (2nd Cir. 1940).
87 Id. at 807.
88 Id. at 809.
89 Recent Decisions, Right of Privacy—Public Concern in Private Life of a Former Public Figure, 18 N.Y.U. L.Q. REV. 133, 133-34 (1940); see also Recent Cases, Right of Privacy—Protection Against the Publication of Newsworthy Information, 2 WASH. & LEE L. REV. 133 (1940); Recent Cases, Torts—Right of Privacy—Public Figure Test as Determinative of Right to Recovery, 8 U. CHI. L. REV. 382 (1941).
90 Warren | Edison | Pavesich | instant | England | principal | Brandeis | Atkinson | Doherty
91 80 N.W. 285 (Mich. 1899) (recognizing no right of privacy in the use of the name and likeness of a deceased person as a label used in the sale of a cigar named after him).
92 67 A. 392 (N.J. Ch. 1907) (recognizing a right of privacy in the unauthorized use of a famous person’s name to promote a product); see, e.g., Note, Right of Privacy—Public Figure—Unauthorized Use of Photograph for Commercial Purposes, 16 TUL. L. REV. 639 (1942) (holding no right of privacy for public figure in use of photograph; comparing, inter alia, Atkinson, with, inter alia, Pavesich); Recent Cases, Photographs—Implied Contract Between Photographer and Customer—Property in Negative—Right of Privacy, 10 U. CIN. L. REV. 336 (1936) (discussing principally Bennett v. Gusdorf, 53 P.2d 91 (Mont. 1936); comparing, inter alia, Atkinson, with, inter alia, Pavesich, and Edison); Recent Cases, Right of Privacy—Nature and Extent—Biographical Motion Pictures, 44 HARV. L. REV. 1146 (1931) [hereinafter Biographical Motion Pictures,] (discussing principally Melvin v. Reid, 297 P. 91 (Cal. Dist. Ct. App. 1931) (holding release of an unauthorized biographical motion picture of plaintiff’s past life of shame violated her right to privacy); citing, inter alia, Pavesich); Recent Cases, Torts—Right of Privacy—Right in the Family, 5 U. CIN. L. REV. 243 (1931) (discussing principally Bazemore v. Savannah
low weighted topic,\textsuperscript{93} is an individual’s right to privacy in broadcasts, which began to be addressed in the 1930s.\textsuperscript{94} In conjunction with the commercial use of a person’s name or likeness, one topic,\textsuperscript{95} identified from relatively few articles, addresses unfair competition.\textsuperscript{96}

Another low weighted topic\textsuperscript{97} is associated with property rights in letters and other literary property—specifically, whether the right to privacy could prevent the unauthorized publication of private letters or other writings.\textsuperscript{98} Finally, wiretapping, the Fourth Amendment, unreasonable searches, and interception of communications were just beginning to be addressed by the late 1930s.\textsuperscript{99}

---


1951 – 1960

The highest weighted topic for this decade is too diffuse to recognize as a “topic,” in and of itself, however the articles from which it is derived reflect commentators’ and courts’ continuing efforts to acknowledge the right to privacy, in some cases extending beyond unauthorized publication of name or likeness. Articles published from 1951 – 1960 still though focused primarily on the unauthorized commercial use of a person’s name or likeness.

The topic models in this decade also reveal discussions, though in very few articles, associated with: debt collection practices; wiretapping; and whether compelling a witness to testify before televised congressional hearings violates the witness’s right of privacy.

---

100 privacy | invasion | person | action | recognized | publication | recognition


103 picture | advertising | commercial | photograph | interest | publicity | unauthorized | photographs | contract

See, e.g., Joseph R. Grodin, Note, The Right of Publicity: A Doctrinal Innovation, 62 YALE L.J. 1123 (1953); Recent Decisions, Television and the Right of Privacy, 3 SYRACUSE L. REV. 208 (1951). In fact, this concept generated a number of topics from the corpus. See, e.g., Burt DeRieux, Comment, Television Tort, 3 MERCER L. REV. 327 (1952); Recent Decisions, Manner of Publication Determinative of Action for Invasion of Privacy, 9 BUFF. L. REV. 362 (1960).

104 debtor | debt | employer | creditor | collection | payment | credit | letter

See, e.g., James Magavern, Recent Decisions, Invasion of Privacy in Debt Collection, 7 BUFF. L. REV. 327 (1958) (discussing Gouldman-Taber Pontiac, Inc. v. Zerbst, 100 S.E.2d 881 (Ga. 1955)); Current Decisions, Torts—Collection Methods—Informing Debtor’s Employer of Employee’s Indebtedness Not an Invasion of Privacy, 26 ROCKY MTN. L. REV. 347 (1954) (discussing Lucas v. Moskins Stores, Inc., 262 S.W.2d 679 (Ky. Ct. App. 1953)); Recent Cases, Torts—Right of Privacy—Malicious Harassment by Collection Agencies, 10 VAND. L. REV. 158 (1956) (discussing Housh v. Peth, 133 N.E.2d 340 (Ohio 1956)). There were a few articles addressing debt collection and privacy in the 1891-1950 corpus, but they were not sufficient enough to result in an identifiable topic. See, e.g., Case Comment, Publication of a Debt—Newspaper’s Liability, 30 KY. L.J. 327 (1942); Recent Cases, Right of Privacy—Infringement—Placarding Debtor as Invasion of the Right, 41 HARV. L. REV. 1070 (1928) [hereinafter, Placarding Debtor].

105 wire | tapping | attorney | general | federal | evidence | security | congress | department | persons | people | justice | laws | communications


106 committee | witness | power | congressional | congress | television | investigation | hearings | questions | purpose | legislative | information | testimony | answer | inquiry

Similar to the previous decade, the highest weighted topics, while reflecting fundamental privacy law issues, were too diffuse to recognize as a specific “topic.” However, the most heavily weighted identifiable topic for this decade again relates to publication of a person’s name or likeness without his or her consent. The second-most heavily weighted identifiable topic arises directly from Griswold v. Connecticut, in which the U.S. Supreme Court recognized a “zone of privacy created by several fundamental constitutional guarantees.”


See, e.g., Eugene N. Aleinikoff, Privacy in Broadcasting, 42 IND. L.J. 373 (1967); Alice G. Donenfeld, Property or Other Rights in the Names, Likenesses or Personalities of Deceased Persons, 16 BULL. COPYRIGHT SOC’Y U.S.A. 17 (1968); Charles Simon, Case Note, Torts—invasion of Privacy—Unauthorized Use of Photograph, 16 DEPAUL L. REV. 255 (1966) (noting that as of yet, no Massachusetts case had established a legally protected right of privacy); Case Note, Right of Privacy—Use of Newsworthy Photograph in Incidental Advertising of Magazine Held No Invasion of Privacy, 31 FORDHAM L. REV. 394 (1962); Comment, Right of Privacy Statute Not Violated by Advertising Designed to Sell Advertising Space, 37 N.Y.U. L. REV. 950 (1962); Recent Decisions, Torts—Privacy Not Abridged When Person’s Name and Photograph Used in Periodical Are Later Used to Advertise Periodical, 14 SYRACUSE L. REV. 136 (1962).


Fourth Amendment searches, specifically through wiretapping, also make an appearance principally due to the Supreme Court’s holding in Katz v. United States. Another relatively weak but identifiable topic primarily addresses privacy in First and Fourteenth Amendment contexts. The fifth-most heavily weighted identifiable topic arises from discussion of Time, Inc. v. Hill, which held that constitutional protections of free speech and the press preclude application of New York’s privacy statute to redress false reports of matters of public interest in the absence of proof that the defendant published the report with knowledge of its falsity or in reckless disregard of the truth, as compared with New York Times v. Sullivan, which limited libel actions under the First Amendment.

The remaining topics, though identifiable, are relatively weak, each reflecting only a few articles within the corpus, addressing: Fourth Amendment searches generally; computers and

---


See David C. Faulkner & Robert I. Friedman, Fourth Amendment Principles and Supreme Court Practice, 25 Ohio St. L.J. 538 (1964); Richard C. Goodman, Comment, Privacy and Political Freedom: Application of the Fourth Amendment to “National Security” Investigations, 17 UCLA L. Rev. 1205 (1970); Note, College Searches and
data collection,\textsuperscript{122} with associated credit collection concerns;\textsuperscript{123} workplace privacy (making its first appearance), primarily reflecting concerns over employee polygraph tests,\textsuperscript{124} and, again, unauthorized use of name or likeness, this time more directly related to imitations and parody.\textsuperscript{125}

**1971 – 1980**

The heaviest-weighted identifiable topic\textsuperscript{126} in this decade reflects generally the control of information about the individual, especially with respect to the government.\textsuperscript{127} The next topic reflects even more clearly concepts of individual information control vis-à-vis the government.\textsuperscript{128} In particular, this decade was a time when scholars and commentators wrote about “data bank privacy”\textsuperscript{129} and first raised concerns about computer matching programs\textsuperscript{130} and digital

\begin{itemize}
\item [122] data | information | computer | system | personal | bureau | access | census | credit | agency | industry | files
\item [123] credit | information | file | access | personal | bureaus
\item [124] employee | lie | employer | detector | test | employment | union | polygraph | board | results | rule
\item [125] copyright | parody | work | unfair | competition | commercial | burlesque | literary | trade | public | light | television | character | copyrighted | motion | defamation | humor | infringement | gas
\item [126] privacy | individual | interests | private | personal | information | protection | control | legal | person | concept | definition
\item [128] information | act | data | records | government | privacy | access | individual | federal | agencies | personal | files
dossiers. The next topic reflects constitutional protection of intimate, personal decisions, first articulated in Griswold v. Connecticut, as extended by not only Roe v. Wade in 1973 but also by Carey v. Population Services International. This topic also reflects the general notions of freedom of personal choice, autonomy, and privacy.

The next two topics both relate to public disclosure of personal facts; one reflecting First Amendment and tort issues associated with disclosures by the press, and the other reflecting disclosures by the government. Two topics reveal issues of electronic surveillance,
eavesdropping, and Fourth Amendment searches,\textsuperscript{140} the second of which also introduces automobile searches,\textsuperscript{141} as well as the third-party doctrine expressed in \textit{United States v. Miller},\textsuperscript{142} which held that the Fourth Amendment does not require the government to obtain a warrant to seize bank records.\textsuperscript{143} Concepts associated with the disclosure and retention of arrest records were identified in a topic,\textsuperscript{144} as well as a topic\textsuperscript{145} reflecting defamation and privacy under the

---


\textsuperscript{140} surveillance | electronic | section | title | eavesdropping | evidence | communications | united | states | conversations | telephone | federal | interception | order | III | act | wiretapping

amendment | fourth | search | evidence | seizure | property | warrant | government | papers | rights | unreasonable


\textsuperscript{142} 425 U.S. 435 (1976)

\textsuperscript{143} \textit{Id.} at 443 (“[I]nformation revealed to a third-party and conveyed by him to Government authorities, even if the information is revealed on the assumption that it will be used only for a limited purpose and the confidence placed in the third party will not be betrayed.”). \textit{Miller’s} holding was limited by the Right to Financial Privacy Act of 1978, Pub. L. No. 95-630, 92 Stat. 3697 (1978) (codified as amended at 12 U.S.C. §§ 3401 – 3422 (2012)), by, for example, requiring the Government authority to notify the bank customer of the subpoena or summons served on the financial institution as well as the nature of the law enforcement inquiry to which the subpoena or summons relates. \textit{See also} Susan Orzack Posen, Comment, \textit{A Paper Chase: The Search and Seizure of Personal Business Records}, 43 BROOK. L. REV. 489 (1977); Note, \textit{Government Access to Bank Records}, 83 YALE L.J. 1439 (1974).


\textsuperscript{144} arrest | records | criminal | police | individual | enforcement | state | dissemination | crime | conviction | justice | agencies | system | juvenile | FBI | arrested | retention | arrestee


\textsuperscript{145} public | defamation | damages | media | standard | liability | false | Gertz | private | interest | amendment | privilege

21
First Amendment, particularly by extending the discussion of *New York Times v. Sullivan*\(^{146}\) and *Gertz v. Robert Welch, Inc.*\(^{147}\)

Additional weaker, and sometimes more diffuse, topics for this decade include: a mixture of themes, including privacy of student records, desegregation, rights of association, and zoning;\(^{148}\) the interplay between privacy and credit reporting;\(^{149}\) with particular emphasis on the Fair Credit Reporting Act;\(^{150}\) medical decisional privacy and autonomy\(^{151}\) related to the right to refuse medical treatment,\(^{152}\) minors’ rights to abortion,\(^{153}\) and disclosure of health data;\(^{154}\) early discussions regarding sexual autonomy;\(^{155}\) witness testimony and immunity;\(^{156}\) and topics

\(^{146}\) 376 U.S. 254 (1964).


\(^{148}\) family | child | parent | association | children | state | intimate | marriage | freedom | parental | relationship | Moore | choice


\(^{149}\) information | consumer | credit | reporting | report | reports | insurance | act | agency | individual | bureau | FCRA | business | investigative


\(^{151}\) medical | patient | treatment | state | physician | consent | health | life | drug | care | relationship


\(^{153}\) moral | constitutional | sexual | human | principles | life | homosexual | morality | rights | love | homosexuality | acts | people | autonomy | theory | values | general

revealing the interplay between copyright and privacy,\textsuperscript{157} issues associated with commercial speech and privacy,\textsuperscript{158} and the right of publicity.\textsuperscript{159}

\textbf{1981 – 1990}

The heaviest weighted substantive topic\textsuperscript{160} in this decade relates to information privacy, reflecting issues associated with the computerized collection and storage of information, both

\textsuperscript{156} privilege | amendment | evidence | witness | compelled | criminal | immunity | information |

\textsuperscript{157} work | author | copyright | protection | rights | property | personal | common | doctrine | contract | letter | trade | unpublished | public


\textsuperscript{158} speech | amendment | commercial | mail | advertising | public | lists | telephone | home | communication | interests | calls | direct | regulation | receive | residential


\textsuperscript{159} publicity | Zacchini | act | commercial | interest | property | appropriation | news | likeness | celebrity | performance | amendment | public | performer | media | entire | interests | rights


A heavier weighted topic reflects privacy more generically:

privacy | tort | action | invasion | private | intrusion | public | person | common | Prosser | facts | Warren

privately and by the government.\(^{161}\) The next most heavily weighted topic\(^{162}\) reflects Fourth Amendment searches.\(^{163}\) The next topic\(^{164}\) reflects almost exclusively discussion of The Florida Star v. B.J.F.,\(^{165}\) which held that a Florida newspaper reporting the name of a rape victim, which had been obtained from public records, in violation of a Florida statute and the newspaper’s own internal policy was protected by the First Amendment.\(^{166}\) The fourth most heavily weighted substantive topic\(^{167}\) identifies concepts associated with false light, defamation, and privacy.\(^{168}\)

The next topic\(^{169}\) reveals concepts associated with privilege,\(^{170}\) specifically related to family testimonial\(^{171}\) and rape victim privileges.\(^{172}\) Medical confidentiality has its own identified


\(^{162}\) amendment | fourth | privacy | search | government | expectation | surveillance | Katz | reasonable | property | private | analysis


\(^{164}\) public | information | media | press | victim | rape | trial | interest | amendment | publication | criminal | access | news | Florida | proceedings | statute

\(^{165}\) 491 U.S. 524 (1989).


\(^{167}\) false | public | light | defamation | speech | private | amendment | standard | malice | publication | actual | truth | libel | actions | defamatory | statement


\(^{169}\) privilege | child | communications | parent | attorney | client | confidential | relationship | disclosure
topic, including AIDS-related medical records. Additional topics include bank and business records privacy, as well as concepts associated with sexual privacy, including those associated with reproductive autonomy.

The last six identified topics for this decade reflect: an extension of the original unauthorized commercial appropriation action, the right of publicity; employee drug and

---

polygraph testing;\textsuperscript{179} various concepts associated with medical decisions,\textsuperscript{180} including sustaining life and euthanasia,\textsuperscript{181} as well as prisoner hunger strikes;\textsuperscript{182} the interplay between privacy and copyright;\textsuperscript{183} emerging technological innovations such as two-way cable\textsuperscript{184} and Caller ID;\textsuperscript{185} and residential privacy, particularly in regard to picketing.\textsuperscript{186}


\textsuperscript{179} employee | testing | drug | employer | test | employment | public | polygraph | private | reasonable | intrusion | job | results

\textsuperscript{180} treatment | patient | life | state | medical | death | suicide | consent | refuse | hunger | prison | care | feeding | punishment


\textsuperscript{184} cable | electronic | telephone | television | communications | service | caller | state | subscriber | information | regulation | FCC | identification | public | commission | content


\textsuperscript{186} speech | subliminal | public | ordinance | picketing | residential | expression | amendment | content | communication
The two most heavily weighted identifiable topics in this decade reflect a wide range of diverse privacy issues. The first identifiable topic relates primarily to government access to and disclosure of personal, financial, and medical data. The next topic addresses the conflict


between the First Amendment freedoms of speech and the press and an individual’s right to privacy, as well as false light invasion of privacy. Closely related, the next topic addresses public disclosure of private facts, primarily with respect to rape victims (continuing the Florida Star discussion), though with some general discussion of regarding publicizing trial proceedings. The next weighted topic identifies again Fourth Amendment searches.
The next three topics, in descending order of weight, identify: information privacy issues associated with the Internet and cyberspace; video surveillance, voyeurism, and thermal imaging; and workplace surveillance and employee privacy, particularly with respect to email.

The next weighted topic again identifies privacy issues associated with reproductive

---


---

abortion | state | women | child | life | children | family | parents | Roe | treatment | pregnancy | decision | pregnant | mother | consent | fetus | birth
rights. Privacy and free speech is again identified as a topic, as are issues associated with genetics and HIV, as well as the right of publicity. Sex offender registration is mixed with a few articles on drivers’ privacy protection, followed closely by a topic identifying issues raised by *Bowers v. Hardwick*. One fairly weak, diffuse topic identifies a range of

---


205 speech | government | amendment | public | free | content | regulation | doctrine | expression | restrictions | audience | interest | message | people | home | justice | picketing


206 health | medical | HIV | care | patient | information | testing | AIDS | physician | disclosure | patients | confidentiality


207 publicity | commercial | identity | celebrity | likeness | White | common | appropriation | copyright | voice | property | public


208 sex | notification | offenders | community | states | registration | public | crime | information | criminal | convicted | released


209 sexual | sodomy | homosexual | sex | Hardwick | Bowers | marriage | conduct | men | laws
“secrecy” issues: executive authority,212 trade secrets,213 and encryption.214 A final collection of lower weighted topics for this decade includes: drug testing generally,215 cell phone privacy and caller ID,216 public disclosure of private facts from the perspective of outing homosexuals,217 and copyright and privacy in unpublished works.218

2001 – 2013

This date range reflects the largest individual collection of documents.219 While the most heavily weighted topic220 in this collection is too diffuse to identify a specific privacy concept


214 drug | testing


215 telephone | caller | ID | privacy | call | service | state | number | communications | cellular | phone | blocking | information | company | wiretap


216 gay | outing | sexual | orientation | people | homosexuality | men | lesbians | sex | homosexual


217 copyright | fair | works | unpublished | work | public


218 A total of 842 published articles.
through its terms alone, its underlying articles reflect a range of modern information privacy issues.\textsuperscript{221} Similarly, the second highest-weighted topic\textsuperscript{222} reflects a wide range of privacy issues.\textsuperscript{223} The next two identifiable topics reflect information privacy concerns, first with a bit of a slant towards consumer privacy\textsuperscript{224} and, second, more generally,\textsuperscript{225} False light invasion of
privacy is a heavily weighted topic.\textsuperscript{226} followed closely by Fourth Amendment searches.\textsuperscript{227} And once again, privacy versus First Amendment rights is a heavily weighted topic.\textsuperscript{228} Computer surveillance, particularly through email, is the next weighted topic,\textsuperscript{229} closely followed by


\textsuperscript{226} tort | false | light | damages | invasion | defamation | Prosser | liability | harm | claim | action | private | facts | disclosure


\textsuperscript{227} amendment | fourth | search | government | police | reasonable | expectation | Katz | warrant | justice | home | enforcement | test | doctrine | evidence | property | supreme


\textsuperscript{228} public | news | press | media | private | life | newsworthy | publication | figures | newsworthiness


\textsuperscript{229} communications | mail | electronic | computer | act | information | access | internet | telephone | wiretap | messages | interception | stored | ECPA | storage | service | system

privacy on the Internet, particularly through social media. Surveillance with respect to national security is a distinctly identified topic, while Bartnicki v. Vopper warrants its own topic with respect to private conversations and the media’s First Amendment rights. Government intrusion upon intimate decisions continues to be identified as a topic, particularly now in regard to the holding in Lawrence v. Texas. The next topic, while addressing information


532 U.S. 514 (2001) (upholding Pennsylvania and federal wiretap acts’ prohibitions against intentional disclosure of illegally intercepted communication which disclosing party knows or should know was illegally obtained as content-neutral laws of general applicability, while holding application of acts against media defendants violated their free speech rights since taped intercepted conversation concerned matter of public importance and defendants had played no part in the illegal interception).


Lawrence | due | marriage | sex | government | supreme | amendment | decision | sexual | interest

539 U.S. 558 (2003) (holding Texas statute making it a crime for two persons of the same sex to engage in certain intimate sexual conduct was unconstitutional, as applied to adult males who had engaged in consensual act of sodomy in privacy of home), overruling Bowers v. Hardwick, 478 U.S. 186 (1986); see, e.g., Sonu Bedi,
privacy, is fairly diffuse, though principally identifying issues associated with access to government records.\textsuperscript{238} Similarly, the next weighted topic\textsuperscript{239} addresses surveillance through a range of modern technologies, including payment systems, autonomous automobiles, ubiquitous monitoring, RFID, GPS, and drones.\textsuperscript{240} The next four topics, in descending order of weight, identify issues associated with: health records privacy, particularly with respect to the Health Insurance Portability and Accountability Act (HIPAA);\textsuperscript{241} workplace monitoring and employee privacy;\textsuperscript{242} family privacy;\textsuperscript{243} and copyright, fair use, and privacy.\textsuperscript{244} The least weighted topics


\textsuperscript{237} information | records | public | access | government | disclosure | privacy | interest | criminal | documents | FOIA | personal | trial | exemption


\textsuperscript{239} technology | surveillance | privacy | information | technologies | GPS | data | tracking | RFID | location | devices | vehicle | public | enforcement | phone | time


\textsuperscript{242} employee | employer | employment | monitoring | workplace | work | reasonable | conduct | surveillance | expectation | business | duty | company | policy | private

for this time period (in descending order) generally reflect a central theme but also represent a diffuse collection of articles: privacy and free speech;\textsuperscript{245} tax privacy;\textsuperscript{246} the right of publicity;\textsuperscript{247} and genetic privacy.\textsuperscript{248}

---


\textsuperscript{243} child | children | family | parents | parental | birth | state | gender


\textsuperscript{244} property | rights | copyright | fair | work | intellectual | email | common | author | contract | expression | owner | interest


\textsuperscript{245} speech | public | expression | amendment | content | free | justice | interest | government


\textsuperscript{246} tax | privacy | information | government | individual | public | disclosure


\textsuperscript{247} publicity | identity | commercial | rights | celebrity | image | property | likeness | trademark | appropriation | claim | athletes | person | individual | economic | persona | advertising


\textsuperscript{248} genetic | DNA | information | testing | individual | samples | identity | crime | discrimination

As noted previously, due to the size of this corpus, we performed an additional 50-topic run in order to potentially identify additional topics that may have been missed in a run limited to twenty-five topics. We found only two additional fairly low weighted topics that were not previously explicitly identified: online consumer tracking, and identity theft (though mixed in with national identification systems).

1891 – 2013

With the exception of including Warren and Brandeis, the four most heavily weighted topics for the entire corpus reflect only general concepts—versus privacy-related concepts. And a heavily weighted topic reflects an interplay between privacy and sexual autonomy; again, not so much directly from its terms as in the documents from which it is derived. From this point forward, in descending order of weight, the topics contain well-recognized terms, and reflect well-recognized concepts, associated with privacy:

- Unauthorized publication of an individual’s picture or likeness;

---

See supra note 68.

See, e.g., Bergerson, supra note 224; Susan E. Gindin, Nobody Reads Your Privacy Policy or Online Contract? Lessons Learned and Questions Raised by the FTC’s Action against Sears, 8 NW. J. TECH. & INTELL. PROP. 1 (2009); Laura J. Bowman, Note, Pulling Back the Curtain: Online Consumer Tracking, 7 I/S: J. L. & POL’Y FOR INFO. SOC’Y 721 (2012); Schmierer, supra note 224.


See, e.g., Sobel, Demeaning of Political Identity, supra note 231; Sobel, Demeaning of Identity, supra note 231.


See, e.g., Warren & Brandeis, supra note 10; Comments on Recent Cases, Torts—Right of Privacy—Nature and Extent—Radio Broadcasting, 25 IOWA L. REV. 387 (1940); Recent Cases, Biographical Motion Pictures, supra note 92; Recent Cases, Constitutional Law—Privacy—Pictures, 19 YALE L. J. 130 (1909) (discussing Foster-Milburn Co. v. Chinn, 120 S.W. 364 (Ky. Ct. App. 1909) (holding publication of the picture of a person without his consent as an advertisement of the publisher’s business is a violation of the right of privacy)); Recent Cases, Torts—Right of Privacy—Publication of Picture of Deceased Child as Invasion of Parents’ Right of Privacy, 15 MINN. L. REV. 610
• False light invasion of privacy;
• Records privacy;
• Fourth Amendment searches;

(1931); Recent Cases, Unauthorized Use of Portrait for Advertising Purposes, supra note 78 (discussing Pavesich v. New England Life Ins. Co., 50 S.E. 68, 77-81 (Ga. 1905)); see also Case Comment, Placarding Debtor, supra note 104.


information | records | government | act | disclosure | access | individual | public | agency | federal | bank | financial | privacy | credit | data | computer


amendment | fourth | search | government | police | warrant | evidence | seizure | justice | property | probable | officers | enforcement


A second, lower weighted, Fourth Amendment topic, more closely associated with technology, was also identified:

fourth | amendment | privacy | expectation | government | reasonable | Katz | search | home | police | enforcement | justice

• Information privacy, generally, along with data mining, as well as more closely related to consumer protection in a separate topic;

• Public disclosure of private facts, particularly with respect to The Florida Star v. B.J.F.;

• Electronic monitoring and surveillance;

• Free speech and privacy, particularly in relation to picketing;

---

261 information | privacy | personal | individuals | control | social | people | public | disclosure


263 information | privacy | data | consumer | personal | companies | industry | FTC | protection | regulation | practices | legislation | collection

264 public | information | press | media | private | amendment | interest | publication | disclosure | facts | news | supreme | speech | matter
See, e.g., 491 U.S. 524 (1989) (holding that imposing damages on newspaper for publishing name of rape victim which had been obtained from publicly released police report, in violation of Florida statute and newspaper’s own internal policy did not comport with First Amendment); see, e.g., Erwin Chemerinsky, Protecting Truthful Speech: Narrowing the Tort of Public Disclosure of Private Facts, 11 CHAP. L. REV. 423 (2008); Deborah W. Denno, Perspectives on Disclosing Rape Victims’ Names, 61 FORDHAM L. REV. 1113 (1993); Ella J. Fenoglio, Note, Poteet v. Roswell Daily Record, Inc.: Balancing First Amendment Free Press Rights Against a Juvenile Victim’s Right to Privacy, 10 N.M. L. REV. 185 (1979-1980); Fishbein, supra note 166; Leone, supra note 196; Stanton, supra note 196; Wey, supra note 196.

266 electronic | communications | mail | telephone | act | interception | computer | service | surveillance | wiretap | monitoring | congress | conversation | messages | ECPA | party
• Workplace monitoring and employee privacy;\(^{268}\)
• Privacy related to minors and sexual offenders;\(^{269}\)
• Privileged communications;\(^{270}\)
• National security surveillance;\(^{271}\)
• Video voyeurism, with some mixture of paparazzi and undercover video;\(^{272}\)
• Health-related privacy, including right-to-die, genetic discrimination, and HIV;\(^{273}\)


\(^{268}\) employee | employer | employment | privacy | testing | drug | workplace | monitoring | private | work | business | test | job | public | reasonable | policy

See, e.g., Chadwick, supra note 242; Hallinan, supra note 202; Noyce, supra note 242; Pagnattaro, supra note 242; Sprague, supra note 242; Joan Vogel, Manufacturing Solidarity: Adventure Training for Managers, 19 HOFSTRA L. REV. 657 (1991); Charles J. Dangelo, Comment, The Individual Worker and Drug Testing: Tort Actions for Defamation, Emotional Distress and Invasion of Privacy, 28 DUQ. L. REV. 545 (1990); Landis, supra note 179; Sipherd & Volpe, supra note 242.

\(^{269}\) child | children | parents | family | state | parental | minor | women


\(^{270}\) privilege | information | disclosure | confidentiality | confidential | relationship | communications | client | attorney | confidence | evidence | discovery | testimony


\(^{271}\) states | united | security | national | surveillance | government | intelligence | act


\(^{272}\) public | intrusion | video | person | images | privacy | reasonable

See, e.g., Calvert & Brown, supra note 201; Christina M. Locke & Kara Carnley Murrhee, Is Driving with the Intent to Gather News a Crime? The Chilling Effects of California’s Anti-Paparazzi Legislation, 31 LOY. L.A. ENT. L. REV. 83 (2011); Pyk, supra note 191; Leadstrom, supra note 201; Mills, supra note 201; Rosenfeld, supra note 228; Lisa F. Wu, Peeping Tom Crimes, 28 PAC. L.J. 705 (1997).
• Fair use of unpublished materials;  
• Online privacy;  
• Right of publicity, particularly related to celebrity imitations.

The 50-topic runs did not identify any topics that had not been identified in earlier corpora.

C. Our Ontology of Privacy Law

The topic modeling employed in this research reveals that privacy law in the United States fundamentally addresses government interference with intimate decisions and protecting private matters from surveillance and exposure. This may appear overly simplistic, but it is not; the complexity of U.S. privacy law lies in degrees—particularly of private matters and surveillance.

---

274 health | medical | genetic | DNA | HIV | individual | research | information | insurance | testing | blood


275 property | copyright | rights | work | fair | author | protection | intellectual | public | unpublished | publication | common | act


276 internet | users | online | user | social | information | web | computer | access | site | digital | Facebook | content | website | cyberspace | Google


277 Incidentally (or perhaps not so incidentally), these privacy protections apply only to corporeal individuals. Despite granting corporations near absolute political free speech rights (see Citizens United v. FCC, 558 U.S. 310 (2010)), as well as freedom of religion privileges (see Burwell v. Hobby Lobby Stores, Inc., 134 S. Ct. 2751 (2014)), the U.S. Supreme Court has held that corporations are not individuals so far as the Freedom of Information Act is concerned. See FCC v. AT&T Inc., 562 U.S. 397 (2011) (holding “personal privacy,” within the meaning of Freedom of Information Act Exemption 7(C), which covers law enforcement records, the disclosure of which could reasonably be expected to constitute an unwarranted invasion of personal privacy, does not encompass corporations).
Beginning with *Griswold v. Connecticut*, and continuing with *Roe v. Wade*, *Carey v. Population Services International*, and *Lawrence v. Texas*, the U.S. Supreme Court has established Fourteenth Amendment privacy protection for “personal decisions relating to marriage, procreation, contraception, family relationships, child rearing, and education. . . . These matters, involving the most intimate and personal choices a person may make in a lifetime, choices central to personal dignity and autonomy, are central to the liberty protected by the Fourteenth Amendment.”

We see that the “original” privacy violation—unauthorized commercial use of a person’s name or likeness, a tort also known as the right of publicity—continues as an active privacy topic through 2013. Likewise, we saw that false light defamation and Fourth Amendment searches remained topics of concern from 1891 through 2013.

Many of the topics we discovered dealt, though, with two main issues:

1. some degree of surveillance or observation of
2. matters which individuals would prefer to remain private.

The devil is in the details of how that surveillance or observation changes over time—whether by wiretapping, GPS devices, or online tracking—and what society considers a “protectable” private matter. We return to MacCormick in that while we may know what privacy is, the challenge is the extent and contours to which we may have a right to privacy.

**D. Limitations and Future Research**

The most obvious limitation to our project is the fact that we built our corpora solely from articles citing to *The Right to Privacy*. As we discussed above, we did this in order to quickly identify a large number of relevant documents; obviously there are many more law review and journal articles that substantively discuss privacy law in the United States that do not cite to *The Right to Privacy*. We also chose to limit the scope of our corpora to articles discussing U.S. privacy law. We certainly recognize there is a large body of scholarship on privacy law issues outside the United States, particularly with respect to Articles 7 and 8 of the Charter of Fundamental Rights of the European Union and the European Community’s Data Protection Directive.

---

278 381 U.S. 497 (1965).
284 *Supra* note 39.
286 *See supra* notes 50-53 and accompanying text.
Protection Directive.\textsuperscript{288} In order to limit our document selection and conversion to a manageable size, we did not include articles undertaking a comparative study of U.S. and European privacy laws (even if the articles cited to The Right to Privacy). We also excluded from our corpora articles that acknowledged an historical privacy-law origin (for example, false light defamation) but which then dealt almost exclusively with a non-privacy topic (such as First Amendment or copyright).

Our results could also be limited by our methodology. As noted above,\textsuperscript{289} there is no consensus as to the most accurate number of topics to discover through the topic modeling process. Although we believe we discovered the most relevant topics by limiting the number of topics to twenty-five, other researchers may discover additional relevant topics using a different number of topics to discover. Finally, our results may have been different if we had used a different topic model software implementation.

Expanding the corpus also could enhance this research. Incorporating articles that do not cite to The Right to Privacy would provide a richer, though more difficult to identify, body of documents, as would articles addressing privacy law outside the United States. The same methodologies could also be applied to case law, both U.S. and international.

IV. CONCLUSION

This paper has presented results from applying latent Dirichlet allocation probabilistic topic modeling algorithms to a document collection comprised of published law review and journal articles from 1891 through 2013—all citing to Warren’s and Brandeis’s The Right to Privacy and substantively discussing privacy law issues. From this process we have developed an ontology that confirms U.S. privacy law protects an individual’s right to be secure from invasion by the government or public into matters of a private nature. The challenge of scholars, policy makers, and courts is to discern the extent of that invasion and the contours of what is considered private.

\textsuperscript{288} Directive 95/46/EC, 1995 O.J. (L 281) 31 (Nov. 23, 1995).

\textsuperscript{289} See supra notes 67-70 and accompanying text.
Appendix A: Conceptual Description of the LDA Algorithm

The LDA algorithm uses Bayesian statistics to discover the latent topics in a corpus of documents. Conceptually, the LDA topic model algorithm uses the following process:

1) The number of topics to identify must be pre-selected. There is no set rule for how many topics are ideal; it depends on the context of the corpus. The topics that are discovered are defined only by the cluster of words associated with them. The researcher may decide to choose a name for each topic that makes sense based on context and experience.

2) The algorithm reads each document in the corpus and repeatedly randomly assigns each word to a topic, so that eventually all the words in the document are probabilistically assigned to the pre-selected number of topics.

3) Two values are calculated for each topic:
   a) \( p(\text{topic } t_1 | \text{ document } d_1) \) = the proportion of words in \( d_1 \) (document1) that are currently assigned to topic \( t_1 \) (topic1). For example, if document \( d_1 \) contains 500 words and 80 of those words are assigned to topic \( t_1 \), then \( p = 0.16 \) (80/500).
   b) \( p(\text{word}_1 | \text{topic } t_1) \) = the proportion of words from the entire corpus assigned to topic \( t_1 \). For example, if there are 10,000 words in the entire corpus, and 350 of those words are assigned to topic \( t_1 \), \( p = 0.035 \) (350/10,000).

The proportions in a) and b) are then multiplied: \( (p(\text{topic } t_1 | \text{ document } d_1) * p(\text{word}_1 | \text{topic } t_1)) = 0.16 * 0.035 = 0.0056 \). After repeating this procedure for each document in the corpus, the process converges on the values that represent topic assignments for each document and for the entire corpus.

Collapsed Gibbs sampling is then used to infer the hidden (latent) thematic structure in the entire population of documents:
   a) The topics in the corpus;
   b) The per-document proportion of topics; and
   c) The per-document, per-word topic assignment.

The 3-level Bayesian model is represented mathematically as:

\[
P(\theta_{1:D}, z_{1:D}, w_{1:D}, \beta, \alpha) = \frac{[(\Gamma(\sum_{i=1}^{k} \alpha_i) / \Pi_{d=1}^{D} \Gamma (\alpha_i) ) * (\theta_1^{(a-1)} \ldots \theta_k^{(a-1)}) * \Pi_{j=1}^{D} \theta_j^{N} ]}{[(\Gamma(\sum_{i=1}^{k} \alpha_i) / \Pi_{d=1}^{D} \Gamma (\alpha_i) ) * \sum_{j=1}^{k} [(\theta_1^{(a-1)} \ldots \theta_k^{(a-1)}) * (\Pi_{n=1}^{N} \sum_{i=1}^{k} \Pi_{j=1}^{V} (\beta_{ij})^w)]]}
\]

The Bayesian equation is a three-level nested probabilistic model.

Because the corpora have large numbers of documents, each containing potentially tens of thousands of words, it would take an extremely long time to calculate all the possible ways of assigning words to topics. The LDA process therefore incorporates a statistical technique known as collapsed Gibbs sampling which samples distributions to obtain approximate identified topics in each document, substantially accelerating the entire process.