What Do We Know About Website Privacy Policies? An Exploratory Study Based on Text Mining

Early stage paper

Yaojie Li
University of New Orleans
yli27@uno.edu

Ying Wang
Northern Illinois University
ywang15@niu.edu

ABSTRACT

Internet users often neglect website privacy policies because of the "transparency paradox" – when the privacy policy languages are lengthy, complex, and granular in details, often requiring sophisticated comprehension. Nevertheless, it is worthwhile for us to identify the critical components and nuance in the privacy policies and understand how companies decide and choose website privacy policies with different concentrations. Drawing upon the institutional theory, we conduct a preliminary data analysis to unveil the main topics and words by leveraging various text mining techniques. Also, we perform cluster analysis to find the vital role of the industrial factor in determining the privacy policy content and theme. Our future research will examine more institutional and organizational factors that can influence companies' online privacy policy-making through a broader dataset.

Keywords

Information privacy, privacy policy, institutional theory, topic modeling, text mining.

INTRODUCTION

"Privacy is paramount to us, in everything we do" that is a recent slogan came from the Google Chrome leadership team (Schuh, 2019), while it can be applied to a broader meaning of "us" and "we," including not only large Internet companies but also other stakeholders who are involved in collecting, storing, and releasing personal information. Indeed, privacy is not a new word in the dictionary of human society since it has historical roots in ancient Greek philosophical discussions, e.g., Aristotle's distinction between the public sphere of political life and the private sphere of domestic life. Notwithstanding, privacy currently contains more novel meaning than before because of the explosion of personal information in the information age and the development of various privacy laws and regulations worldwide.

According to a recent survey by the Pew Research Center (2019), most Americans believe that their personal data is less secure now than it was five years ago, and 81% think the potential risks of companies collecting data about them outweigh the benefits. While most consumers feel concerned about and lack control of their data, they would rather blame companies for losing personal data than malicious attackers and request that companies be proactive about data protection (RSA, 2019). Whereas no general federal or state law requires a company to have a privacy policy in all circumstances, many specific laws are enacted to guide company privacy behaviors. Examples include the Children's Online Privacy Protection Act (COPPA), the Gramm-Leach-Bliley (GLB) Act for financial institutions, the Health Insurance Portability and Accountability Act (HIPAA) in healthcare services, and many states' privacy laws, such as California Privacy Rights Act (CPRA) and Virginia Consumer Data Protection Act (VCDPA). Nevertheless, many companies reserve the express right to change the terms of privacy policies unilaterally, and the privacy policy languages are often too long, complex, and granular in details that require an advanced reading capability to comprehend. While many Internet users have been asked to agree to a privacy policy frequently, few read the terms and conditions or understand them completely (McDonald & Cranor, 2008; Pew Research Center, 2019). For that, "the transparency paradox" occurs when all the details are transparently presented in a privacy policy, yet it is beyond the comprehension of most people (Nissenbaum, 2022).

Therefore, Internet users' concerns and confusion can be further magnified because of the disconnect between their comprehension and the privacy practice, e.g., what type of personal information is involved when individuals engage in online activities on a specific website, which parties are involved in collecting, sharing, and using the information, whether and how individuals' privacy can be protected in accordance with relevant privacy laws and regulations. To that end, natural language processing techniques can be used to unveil the myth of complex policy statements (Wilson et al., 2016). Furthermore, natural language processing techniques can help us better understand which aspects of customer privacy companies (i.e., technology, financial, and healthcare sectors) will concentrate on. To sum up, this research aims to address the following questions: 1) what the main content of website privacy policies is, 2) to which extent are those privacy policies related to relevant laws and regulations, and 3) what factors (i.e., institutional factors) can be attributed to the discrepancy in representing relevant privacy laws and regulations? In this early-stage paper, we attempt to address the first two questions while leaving the last one for future research.

To the best of our knowledge, this paper is among the first attempts to analyze website privacy policies using an unsupervised learning method, though prior efforts are based on manual selection and analysis with a limited sample size and supervised learning algorithms. Also, our study will exceed prior efforts in examining the relationship between corporate privacy policy-making and institutional factors, such as regulatory factors, industrial factors, and corporate reliance on information technology and the Internet. Furthermore, our study will complement prior individual-level studies with organizational and institutional perspectives, thus expanding the landscape of information privacy research.

The remainder of this article is structured as follows. First, we describe the research background by reviewing the institutional theory that guides our future research. Then, we present a pilot study using California-based Fortune 500 companies' website privacy policies and interpret the results. Lastly, we discuss promising avenues for future research.

RESEARCH BACKGROUND

Most organizations face information privacy as a tremendous challenge since data use has dramatically increased due to information technology advances while protecting an individual's privacy preference and personally identifiable information has become prevalent. As a result, this topic has attracted much attention from different research areas, including legal studies, organizational studies, computer science, and information systems. There are many definitions for information privacy, while most stress the importance of an individual's control over the potential secondary uses of his or her personal information – using the data for another purpose other than its original one when collecting the data (Bélanger et al., 2002; Bélanger & Crossler, 2011). Also, Smith and colleagues concluded four dimensions of information privacy: collection, unauthorized secondary use, improper access, and errors. Specifically, Clarke (1999) defined information privacy as "the interest an individual has in controlling or at least significantly influencing, the handling of data about themselves." Despite numerous independent information privacy studies, several important literature reviews and theoretical framework studies (Bélanger & Crosser, 2011; Li, 2011; 2012; Pavlou, 2011; Smith et al., 2011) encapsulate critical theories and constructs, describe the status quo of the research, and shed light on future directions. For example, Bélanger and Crosser (2011) identified various information privacy topics and constructs, discussed the theoretical contributions of those studies based on Gregor's information systems theoretical taxonomy, and concluded with an information privacy concern multilevel framework. Smith, Dinev, and Xu (2011) adopted another way to classify the information privacy literature and identified three main areas: "the conceptualization of information privacy, the relationships between information privacy and other constructs, and the contextual nature of these relationships." Lastly, Smith and colleagues recommended an overarching macro model based on antecedents, privacy concerns, and outcomes.

In contemporary society, information has become an essential part of our everyday life, and organizations must adapt to the ever-changing environment to protect the privacy of their customers. While each organization has its own corporate and website privacy policy, its responses to information security and privacy problems have been profoundly influenced by the social, political, economic, and legal forces of the environment where they operate. These institutional environments are characterized by the elaboration of rules and requirements that each organization must conform to gain acceptance and legitimacy in their "institutional field" (Scott, 1995; Suchman, 1995; Lawrence et al., 2002). According to the institutional perspective, organizations are "suspended in a web of values, norms, beliefs, and taken-for-granted assumptions" (Barley & Tolbert, 1997). This perspective has been widely used in examining the internal and external influences on various organizational patterns and explaining why specific organizational structures can survive in the long term (Weerakkody et al., 2009). Specifically, institutions exert three types of pressure on organizations: coercive, normative, and mimetic (DiMaggio & Powell, 1983). Because of these pressures, an organization has to adopt a value system accepted by other organizations in their field as legitimate, and this process is termed institutional isomorphism. Specifically, coercive pressure stem from political power such as governmental policies and regulations, mimetic pressure occurs when there is a need to imitate successful model from competitors to address environmental uncertainties and ambiguities, particularly when there is little understanding about a new policy, process, or technology, and normative pressure arises from the norms embedded in the professionals (DiMaggio & Powell, 1983; Guler et al., 2002).

In addition to sociology and organizational studies where institutional theory emerges and prevails, a large number of Information Systems studies have adopted this theoretical lens to examine IT-related phenomena in organizational settings, such as IT innovation (Geels, 2004), IT development and implementation (Liang et al., 2007), and IT adoption and use (Teo et al., 2003; Zheng et al., 2013), and IT security (Hu et al., 2007; Hsu et al., 2012; Cavusoglu et al., 2015). In contrast, few studies extend the institutional theory to the context of information privacy. It can be explained that prior IS privacy studies focused on individual levels, such as information privacy concerns, information privacy and e-business impacts, information privacy attitudes, information privacy practices, information privacy tools and technologies (Bélanger & Crosser, 2011). Nevertheless, as more individuals acknowledge the importance of their information privacy when interacting with various websites and online platforms, it is paramount to examine organizational responses to information privacy concerns. Furthermore, companies must formulate strategies to address security and privacy issues to obtain “legitimacy” in the new institutional field through coercive, normative, and mimetic isomorphism. Therefore, examining how those companies behave toward a more information privacy-oriented institutionalization under various environmental conditions is also worthwhile.

A PILOT STUDY USING CALIFORNIA FORTUNE 500 COMPANIES

To investigate how companies comply with data privacy regulations, we conducted a pilot study by analyzing privacy policies posted by California-based Fortune 500 corporations on their websites and comparing them with California Consumer Privacy Act (CCPA, 2018) and California Business and Professional Code – Internet Privacy Requirements (2013). Our data set contained 126 companies headquartered in California and on the Fortune 500 list, of which one company did not post a privacy policy on its website. Therefore, the total sample size was 125. We first applied the Latent Dirichlet Allocation (LDA) technique to extract topics from privacy policies and regulation files. Then we conducted a clustering analysis to group companies based on the topic distributions. The level of compliance was measured by the content similarity between organizational privacy policies and governmental rules.

We conduct text analysis in Python 3.9.5. We started text analysis by reading the text file and decomposing documents, a process that decomposes sentences into a bag of words. Then, we conducted a series of pre-processing activities to clean the data in a basic manner, including part of speech (POS) tagging, removing stop words, lowering characters, removing numbers and punctuation, and stemming. Two types of stop words were removed: (1) default English words like “am,” “this,” and “the;” and (2) the unique terms that appear in only one document by following the same procedure proposed in prior studies (Sidorova et al., 2008). After data cleaning, the original textual sentences were converted into a term-document matrix, demonstrating the frequency of each term appearing in each document. Further text analytics was conducted based on this matrix.

We applied Latent Dirichlet Allocation (LDA) to extract topics embedded in privacy policies and regulations. LDA is a text mining technique based on a generative probabilistic model for collections of discrete data (Blei et al., 2003). LDA represents documents that refer to privacy policies and regulations as mixtures of topics that spit out words with specific probabilities. It assumes that documents are a mixture of topics, and topics are a mixture of words. From a technical perspective, LDA is an extension of probabilistic latent semantic analysis (pLSA) introduced by (Hofmann, 2013). LDA assumes that a document d is generated with a probability $p(d)$, and a topic z is picked with a probability $p(z/d)$, and a word w is generated with a probability $p(w/z)$. The probability of finding a word in a document is given by:

$$P(w_i) = \sum_{j=1}^{T} P(w_i/z_{i=j})P(z_{i=j})$$

where $P(z_{i=j})$ is the probability that the topic j for the i^{th} word and $P(w_i/z_{i=j})$ is the conditional probability that the word w_i given the fact that topic j was selected for the i^{th} word. The critical mixture probabilities follow the Dirichlet multinomial distribution, which is given below:

$$Dir(\alpha_1, \dots, \alpha_T) = P(P_1, P_2, \dots, P_T/\alpha_i, \dots, \alpha_T) = \frac{\Gamma \sum_j \alpha_j}{\prod_j \Gamma \alpha_j} \prod_{j=1}^{T} P_j^{\alpha_{j-1}}$$

With unknown parameters $\alpha_1, \dots, \alpha_T$, the formula for the Dirichlet Distribution is a formula that is very difficult to program on a computer. To make things easier, the Dirichlet is used with a constant parameter $\alpha_i = \alpha_j = \alpha$, and Blei et al. (2003) propose that computing the posterior distribution of topics conditioned on given words is an alternative way to obtain the $p(w/\alpha, \beta)$.

The rationale is that it is easier to start with words and determine topics than the other way around. The equation used for the posterior distribution of the topics z conditioned on the words is shown as below:

$$p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)}$$

Blei et al. (2003) also use the above conditional probability to determine the likelihood of selecting words $p(w|\alpha, \beta)$, assuming that both the prior $p(z)$ and the posterior probability $p(w|z)$ followed multinomial distribution from Dirichlet distributions mixtures with parameters (α, β) and concluded that this distribution $p(w|\alpha, \beta)$ that could have been used to generate documents by the formula below:

$$p(w|\alpha, \beta) = \frac{\Gamma \sum_i \alpha_i}{\prod_i \Gamma \alpha_i} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \prod_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{i,j})^{w_n^j} \right) d\theta$$

After extracting the topics from documents, we separately compared topic distributions between each privacy policy and two regulations (California Consumer Privacy Act and Internet Privacy Requirements) to measure the compliance level. For each document, a probability value is assigned to each discovered topic, and the values sum up to 1. In summary, each document is represented by a topic distribution $T_i = \{T_{i,1}, T_{i,2}, \ldots, T_{i,k}\}$ where $T_{i,k}$ is the weight on the K^{th} topic in the document i and $\sum_{k=1}^{K} T_{i,k} = 1$.

We interpreted the discovered topics as content components mentioned by the privacy policy and regulations. If a particular $T_{i,k}$ has the value of 0, it means that topic k is not mentioned in the document i. We define the content similarity $S_i(p,r)$ between privacy policy p and regulation r as the cosine similarity of the two corresponding topic distributions T_p and T_r which can be written as follows:

$$S_i(p,r) = \frac{T_p \cdot T_r}{\|T_p\|\|T_r\|} = \frac{\sum_{k=1}^{K} T_{p,k} T_{r,k}}{\sqrt{\sum_{k=1}^{K} (T_{p,k})^2} \sqrt{\sum_{k=1}^{K} (T_{r,k})^2}}$$

The resulting similarity values are between 0 and 1. If there is no common topic, the numerator will be 0; if the topic distributions of two documents are the same, the similarity will be 1, which is the maximum value. A higher value means a higher content similarity between the privacy policy and regulation, indicating a higher compliance level.

Thus, we created a word cloud to visualize the most critical terms mentioned in privacy policy and regulations. As shown in Figure 1, the words that occurred the most frequently in data privacy documents include “inform,” “service,” “privacy,” “policy,” “data,” “product,” “account,” “site,” “custom,” “provide,” etc.

In the LDA analysis, a standard procedure for evaluating model performance is holding out 20% of data for testing purposes and using the remaining 80% to learn the model (Al-Ramahi, Liu, & El-Gayar, 2017). Considering the small sample size in this pilot study, we applied 5-fold cross-validation to enhance the estimation of model performance. We partitioned the data into five random equal-sized subsamples. A single subsample is retained as validation for testing, and the remaining four are for training. The final result is the average of all test results.

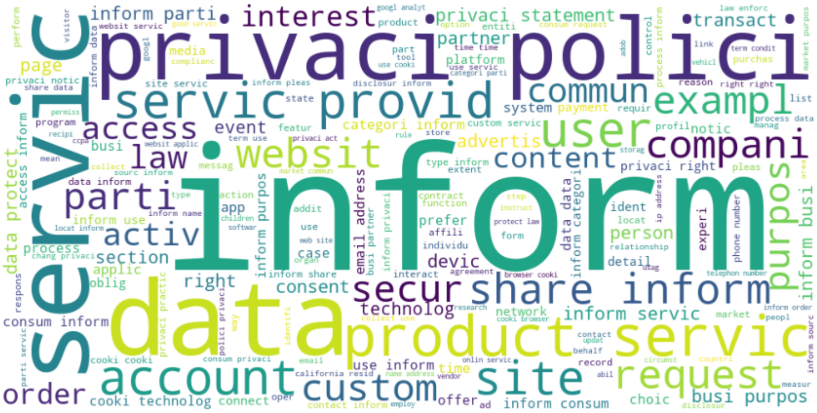

Figure 1. Word Cloud

To evaluate how well a model fits the data, we computed the perplexity of the held-out test set by varying the values of the number of topics, k. Perplexity is a commonly used measurement to evaluate how well a statistical model describes a dataset, with lower perplexity denoting a better probabilistic model (Zhao et al., 2015). As the number of topics increases, perplexity decreases, indicating a better model performance. However, beyond a particular value of k, increasing the number of topics almost has no impact on perplexity, indicating that the model performs best when k equals one specific value without increasing computing complexity. Thus, the value at “the elbow point” is the optimal number of topics. LDA results are terms with high weights in topics and topic probability distributions over textual documents.

We obtained three topics from the LDA analysis and explored topic coherence to justify the model selection. Below are the top words in three topics, and we labeled the topics by combining the meaning of top words and unique words (Table 1). Based on the preliminary results we found in the pilot test, the three topics can be described as commerce, contract, and compliance (Figure 2). The Commerce topic stresses the importance of the commercial relationship and transactions between the company and customers in policy statements, whereas the contract topic accentuates the terms, agreements, and content in website privacy policies. Lastly, the compliance topic concerns customers' privacy rights and how one company complies with those laws and regulations.

Topic	Word	p	Topic	Word	p	Topic	Word	p
0	data	0.032	1	data	0.030	2	**consum**	0.071
0	product	0.013	1	site	0.016	2	**section**	0.021
0	user	0.011	1	cooki	0.013	2	**titl**	0.016
0	websit	0.011	1	websit	0.012		categori	0.013
0	compani	0.009	1	product	0.011	2	**person**	0.013
0	cooki	0.009	1	custom	0.007	2	**subdivis**	0.012
0	devic	0.009	1	address	0.007	2	**pursuant**	0.009
0	access	0.007	1	access	0.007	2	**agenc**	0.008
0	advertis	0.006	1	advertis	0.007	2	**california**	0.007
0	market	0.006	1	email	0.007	2	**regul**	0.006
0	**notic**	0.006	1	process	0.007	2	**act**	0.006
0	custom	0.006	1	devic	0.006	2	**health**	0.006
0	exampl	0.006	1	market	0.006	2	**sale**	0.006
0	technolog	0.005	1	technolog	0.005	2	**code**	0.005
0	address	0.005	1	**pleas**	0.005	2	**state**	0.005
0	**locat**	0.005	1	compani	0.005	2	**action**	0.005
0	commun	0.005	1	browser	0.005	2	**violat**	0.005
0	**program**	0.005	1	**consent**	0.005	2	**identifi**	0.004
0	**payment**	0.005	1	commun	0.005	2	**paragraph**	0.004
0	activ	0.005	1	user	0.005	2	**collect**	0.004
0	process	0.005	1	activ	0.005	2	**ferri**	0.004
0	order	0.004	1	**content**	0.004	2	**disclosur**	0.003
0	email	0.004	1	**protect**	0.004	2	**month**	0.003
0	categori	0.004	1	order	0.004	2	**contract**	0.003
0	**transact**	0.004	1	**contact**	0.004	2	**pg**	0.003
0	**insur**	0.004	1	**name**	0.004	2	**year**	0.003
0	**type**	0.004	1	**statement**	0.004	2	**entiti**	0.003
0	browser	0.004	1	exampl	0.004	2	**behalf**	0.003
0	**affili**	0.004	1	**term**	0.004	2	**direct**	0.003
0	site	0.004	1	**practic**	0.004	2	**attorney**	0.003

Note: Unique words in each topic are shown in boldface.

Table 1. Topics Embedded in Website Privacy Policies

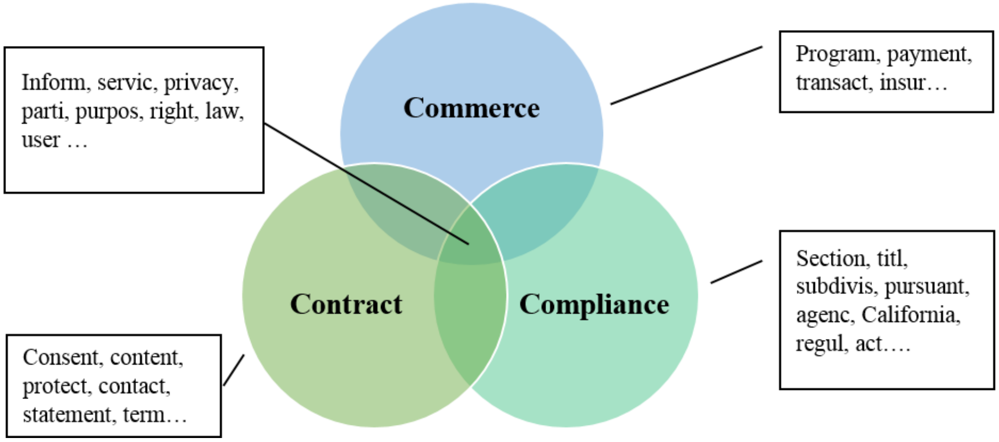
Inform, servic, privacy, parti, purpos, right, law, user …
Commerce
Program, payment, transact, insur…
Contract
Compliance
Section, titl, subdivis, pursuant, agenc, California, regul, act….
Consent, content, protect, contact, statement, term…

Figure 2. Main Topics in Privacy Policies

Also, we conducted a k-means cluster analysis to group companies based on the topic distributions to explore further the relationship between firm characteristics and privacy policy compliance. Since the k-means algorithm requires the number of clusters k to be selected in advance, we calculated the distortion scores for models with a varying value of the number of clusters to determine the optimal value of k. Hence, the best value of k is three (Figure 3), indicating that the firms in our data set could be grouped into three clusters (Figure 4).

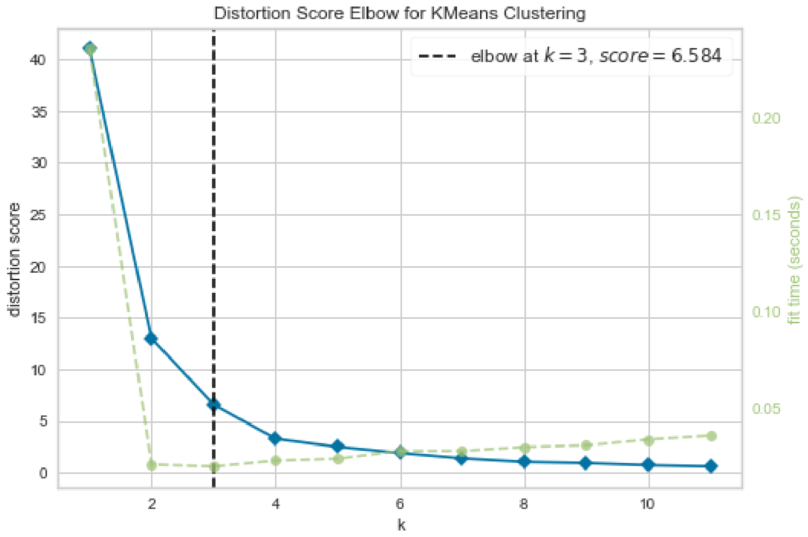

Figure 3. Distortion Score "Elbow" Plot

Table 2 presents a sample of results, showing the LDA analysis, content similarity comparison, and cluster analysis for 15 companies. As shown in Table 3, the distribution of industries and sectors where the California companies are located varies across three clusters. There is a significant difference in the percentages of combined technology and communication services sectors among three groups, 50.75%, 45.45%, and 28.00%, respectively, which suggests possible institutional isomorphism in specific privacy policy-making in hi-tech industries. Compared with Cluster 0 characterized by hi-tech nature (i.e., technology, communication services, healthcare), Cluster 1 and Cluster 2 contain more traditional industrial components (i.e., consumer cyclical and defensive, Industrials, energy, & utilities). Despite these descriptive results, it is worthwhile to examine the influence of institutional and organizational factors on website privacy policy-making using regression analyses in further development.

Company	**Topics**			**Similarity**		**Cluster**
	Commercial Transaction	Contract	Regulatory Compliance	Compliance with CPA	Compliance with CPR	Label
001_Apple	0.999	0.000	0.000	1.000	0.955	1
002_Google	0.680	0.319	0.000	0.905	0.990	1
003_Chevron	0.194	0.805	0.000	0.234	0.511	0
004_Facebook	0.999	0.000	0.000	1.000	0.955	1
005_WFB	0.046	0.952	0.000	0.048	0.341	0
006_Intel	0.284	0.715	0.000	0.370	0.628	0
007_Disney	0.812	0.187	0.000	0.974	0.997	1
008_HP	0.285	0.715	0.000	0.370	0.628	0
009_Cisco	0.990	0.000	0.000	1.000	0.955	1
010_Tesla	0.983	0.017	0.000	1.000	0.960	1
011_Amgen	0.991	0.000	0.000	1.000	0.955	1
012_Netfli	0.801	0.154	0.045	0.981	0.992	1
013_Gilead	0.996	0.000	0.000	1.000	0.955	1
014_TD Synne	0.316	0.323	0.360	0.547	0.688	2
015_Broadcom	0.109	0.411	0.480	0.170	0.352	2

Table 2. Results – Text Mining, Similarity, and Cluster Analysis

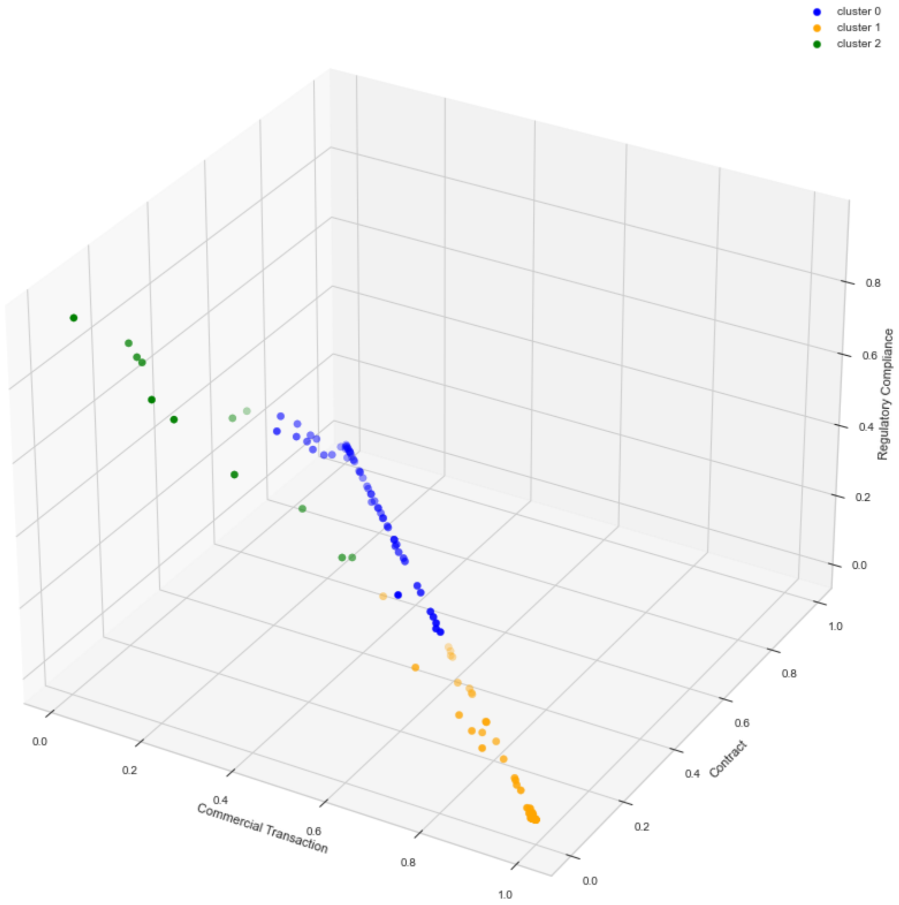

Figure 4. Visualization of Cluster Analysis Results

	Cluster 0		Cluster 1		Cluster 2	
No. of Companies	67	100%	33	100. %	25	100 %
Technology & comm. services	34	50.75%	15	45.45%	7	28.00%
Financial services	8	11.94%	1	3.03%	5	20.00%
Healthcare	9	13.43%	3	9.09%	3	12.00%
Consumer cyclical & defensive	8	11.94%	8	24.24%	6	24.00%
Industrials, energy, & utilities	5	7.46%	4	12.12%	3	12.00%
Others	3	4.48%	2	6.06%	1	4.00%

Table 3. Summary of Cluster Analysis Results

FUTURE RESEARCH AND DISCUSSION

In this early-stage paper, we propose a text mining method to disclose the nature and properties of website privacy policies from an organizational perspective. Also, we attempt to examine how institutional and organizational factors will affect website privacy policy-making. Specifically, we extracted keywords and topics concerning the commercial relationship between companies and customers, the contract content, and companies' compliance with relevant privacy laws and regulations through analyzing 125 California Fortune 500 companies' website privacy policies.

In the following research, we attempt to collect a larger dataset based on the entire Fortune 500 corporations, including California and other states. This can mitigate the possible bias due to a limited sample from one state. In addition, more diversified industries and sectors and local privacy laws and regulations will be included in the further examination. Also, we will test multiple regression models with various institutional variables, such as industrial and sectional categorical variables, technology reliance, and Internet reliance. Last but not least, we will make more effort to provide theoretical and practical contributions based on our empirical inquiry into website privacy policies.

REFERENCES

Al-Ramahi, M.A., Liu, J. and El-Gayar, O.F., 2017. “Discovering design principles for health behavioral change support systems: A text mining approach,” *ACM Transactions on Management Information Systems* (8: 2-3), pp.1-24.

Barley, S.R. and Tolbert, P.S., 1997. “Institutionalization and structuration: Studying the links between action and institution,” *Organization Studies* (18:1), pp.93-117.

Bélanger, F. and Crossler, R.E., 2011. “Privacy in the digital age: a review of information privacy research in information systems,” *MIS Quarterly* (35:4), pp.1017-1041.

Belanger, F., Hiller, J.S. and Smith, W.J., 2002. “Trustworthiness in electronic commerce: the role of privacy, security, and site attributes,” *The Journal of Strategic Information Systems* (11: 3-4), pp.245-270.

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. “Latent Dirichlet allocation,” Journal *of Machine Learning Research* (3), pp.993-1022.

California Business and Professions Code, 2013. Business and profession code, division 8. Special business regulations, chapter 22. Internet privacy requirements. Available from: https://law.justia.com/codes/california/2020/code-bpc/division-8/chapter-22/.

California Consumer Privacy Act. 2018. California Secretary of State. Available from: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375.

Cavusoglu, H., Cavusoglu, H., Son, J.Y. and Benbasat, I., 2015. “Institutional pressures in security management: Direct and indirect influences on organizational investment in information security control resources,” *Information & Management* (52: 4), pp.385-400.

Clarke, R., 1999. “Internet privacy concerns confirm the case for intervention,” *Communications of the ACM* (42: 2), pp.60-67.

DiMaggio, P.J. and Powell, W.W., 1983. “The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields,” *American Sociological Review* (48:2), pp.147-160.

Geels, F.W., 2004. “From sectoral systems of innovation to socio-technical systems: Insights about dynamics and change from sociology and institutional theory,” *Research Policy* (33: 6-7), pp.897-920.

Gregor, S., 2006. “The nature of theory in information systems,” *MIS Quarterly* (30:3), pp.611-642.

Guler, I., Guillén, M.F. and Macpherson, J.M., 2002. “Global competition, institutions, and the diffusion of organizational practices: The international spread of ISO 9000 quality certificates,” *Administrative Science Quarterly* (47:2), pp.207-232.

Hofmann, T., 2013. *Probabilistic Latent Semantic Analysis*. *arXiv preprint arXiv:1301.6705*.

Hsu, C., Lee, J.N. and Straub, D.W., 2012. “Institutional influences on information systems security innovations,” *Information Systems Research* (23:3), pp.918-939.

Hu, Q., Hart, P. and Cooke, D., 2007. “The role of external and internal influences on information systems security–a neo-institutional perspective,” *The Journal of Strategic Information Systems* (16:2), pp.153-172.

Lawrence, T.B., Hardy, C. and Phillips, N., 2002. “Institutional effects of interorganizational collaboration: The emergence of proto-institutions,” *Academy of Management Journal* (45:1), pp.281-290.

Li, Y., 2011. “Empirical studies on online information privacy concerns: Literature review and an integrative framework,” *Communications of the Association for Information Systems* (28:1), pp. 453-496.

Li, Y., 2012. “Theories in online information privacy research: A critical review and an integrated framework,” Decision *Support Systems* (54;1), pp.471-481.

Liang, H., Saraf, N., Hu, Q. and Xue, Y., 2007. “Assimilation of enterprise systems: the effect of institutional pressures and the mediating role of top management,” *MIS Quarterly* (31:1), pp.59-87.

McDonald, A.M. and Cranor, L.F., 2008. “The Cost of Reading Privacy Policies,” I*/S: A Journal of Law and Policy for the Information Society* (4:3), pp. 540-565.

Nissenbaum, H., 2022. “Excerpt from A Contextual Approach to Privacy Online,” *Ethics of Data and Analytics*, pp. 112-118. Auerbach Publications.

Pavlou, P.A., 2011. “State of the information privacy literature: Where are we now and where should we go?” *MIS Quarterly* (35:4), pp.977-988.

Pew Research Center, 2019. Americans and privacy: Concerned, confused, and feeling a lack of control over their personal information. Available from: https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/.

RSA. 2019. RSA data privacy and security survey 2019: The growing data disconnect between consumers and businesses. Available from: https://pdfs.semanticscholar.org/03a4/3636e0b6718c71a936d0237503e9a87719df.pdf.

Scott, W. R. 1995. *Institutions and organizations* (Vol. 2). Thousand Oaks, CA: Sage.

Schuh, J. 2019. Building a more private web. Available from: https://www.blog.google/products/chrome/building-a-more-private-web/.

Sidorova, A., Evangelopoulos, N., Valacich, J.S. and Ramakrishnan, T., 2008. Uncovering the intellectual core of the information systems discipline. *MIS Quarterly* (32:3), pp.467-482.

Smith, H.J., Dinev, T. and Xu, H., 2011. “Information privacy research: an interdisciplinary review,” *MIS Quarterly* (35:4), pp.989-1015.

Suchman, M.C., 1995. “Managing legitimacy: Strategic and institutional approaches,” *Academy of Management Review* (20:3), pp.571-610.

Teo, H.H., Wei, K.K. and Benbasat, I., 2003. “Predicting intention to adopt interorganizational linkages: An institutional perspective,” *MIS Quarterly* (27:1), pp.19-49.

Weerakkody, V., Dwivedi, Y.K. and Irani, Z., 2009. “The diffusion and use of institutional theory: a cross-disciplinary longitudinal literature survey,” *Journal of Information Technology* (24:4), pp.354-368.

Wilson, S., Schaub, F., Dara, A.A., Liu, F., Cherivirala, S., Leon, P.G., Andersen, M.S., Zimmeck, S., Sathyendra, K.M., Russell, N.C. and Norton, T.B., 2016, August. “The creation and analysis of a website privacy policy corpus,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1330-1340.

Zhao, W., Chen, J.J., Perkins, R., Liu, Z., Ge, W., Ding, Y. and Zou, W., 2015. “A heuristic approach to determine an appropriate number of topics in topic modeling,” in *BMC Bioinformatics* (16:13), pp. 1-10. BioMed Central.

Zheng, D., Chen, J., Huang, L. and Zhang, C., 2013. “E-government adoption in public administration organizations: integrating institutional theory perspective and resource-based view,” *European Journal of Information Systems* (22:2), pp.221-234.