GraphDPR: A Privacy Policy Analysis Framework Using Knowledge Graphs and Topic Modeling

Himadri Chowdhury[1], Md Istiak Morsalin, Rafe Sumnan Azade, Vijayalakshmi Ramasamy, and Gokila Dorai[2]

[1] Department of Computer Science, Georgia Southern University, Statesboro, GA, USA
{hc10257, mm58994, ra11045, vramasamy}@georgiasouthern.edu
[2] School of Computer & Cyber Sciences, Augusta University, Augusta, GA, USA
gdorai@augusta.edu

Abstract. Privacy policies play a crucial role in disclosing organizational data practices; however, their lengthy and complex nature hinders user understanding and regulatory auditing, particularly in e-commerce. To address these challenges, we introduce the Data Protection Regulation analysis (GraphDPR) framework, which leverages graph-based semantic analysis for auditing privacy policies. GraphDPR employs transformer-based text processing, knowledge graph creation, and unsupervised topic modeling to generate structured representations of policy content. It converts privacy policies into entity–category–data point triples, normalizes them with Sentence-BERT embeddings, and enhances them into company-specific knowledge graphs using Neo4j. These graphs are then analyzed with Latent Dirichlet Allocation (LDA) to identify thematic patterns in the data collection. GraphDPR facilitates both static and comparative audits by aligning policy content with regulatory standards, yielding interpretable insights into compliance. Experimental results indicate that it provides better regulatory coverage and topic clarity than existing systems, like PolicyGPT and Poligraph. By integrating graph mining and semantic modeling, GraphDPR enhances automated privacy policy auditing and supports scalable compliance monitoring.

Keywords: Privacy Policy Analysis, Knowledge Graphs, Compliance Scoring, Topic Modelling

1 Introduction

E-commerce platforms generate volumes of consumer data, which in turn increases the need for transparent and reliable data practices. Yet, privacy policies (PP) remain difficult for users to interpret due to their legal complexity, inconsistent structure, and ambiguous language [9, 15], which undermines informed consent and complicates the enforcement of regulations like the GDPR and CCPA [3]. Advances in NLP and knowledge graph (KG) modeling have

enabled automated analysis of PPs by extracting structured entities and obligations, using techniques such as sentence classification, summarization, and LLMs like PolicyGPT [1, 6, 8, 14]. However, these methods still face challenges including limited semantic normalization, weak alignment with regulatory obligations, and insufficient comparative analysis across companies and jurisdictions. Even state-of-the-art approaches, such as PolicyGPT [14] and Poligraph [6], often lack interpretability and traceability in regulatory contexts. Our approach bridges technical advances in machine learning (ML) with the demands of regulatory compliance and transparency in the digital economy.

1.1 Problem Formulation/Objectives and Research Questions

We present a multi-stage, graph-based framework **Graph-driven Data Privacy Regulation analysis (GraphDPR)**, for auditing e-commerce PPs and their alignment with GDPR. It enables interpretable, regulation-aware comparisons of privacy disclosures across companies, using a graph mining [10] pipeline, semantic clustering, and knowledge-based reasoning. Using PoliGraph-ER [6] for structured extraction, we transform annotated GraphYML files into quadruples stored in a Neo4j KG. We used Sentence-BERT embeddings [12] for semantic normalization and clustering, and graph paths were linearized for LDA topic modeling [4], enabling interpretable company profiles.

Key contributions include using: (i) GraphDPR, a graph-based framework for regulatory policy interpretation, (ii) semantic normalization to support cross-company comparisons, (iii) interoperable KGs capturing privacy disclosures, and (iv) graph-linearized topic modeling for theme extraction aligned with GDPR.

Unlike prior systems like Poligraph [6], which analyze individual documents, GraphDPR unifies multiple policies into a normalized KG, supporting scalable, comparative, and regulation-conscious analysis of privacy practices, we address the research questions: **RQ1:** *What common user data collection patterns are found across privacy policies of similar companies?* – Identify recurring data types and data collection outliers. **RQ2:** *How are third-party data sharing practices disclosed, and can a unified graph reveal discrepancies?* – Evaluate if consolidated graphs highlight ambiguous or unique sharing disclosures. **RQ3:** *Can graph-linearized topic modeling surface interpretable, regulation-relevant themes?* – Assess if LDA on graph paths provides actionable, GDPR-aligned insights.

2 Literature Review

Privacy policies (PPs) are essential for disclosing data practices, but are often hard to interpret due to complex language and legal jargon. Early research relied on manual methods to assess PPs, but recent studies utilize ML and natural language processing (NLP) for better transparency and compliance. [1, 5, 15]. Supervised ML, such as transformer-based models, allows fine-grained classification of policy content [1]. Reviews highlight persistent challenges of unclear terminology and lack of standardized evaluation [15]. Semi-automated summarization combines NLP and expert input to identify high-risk statements. [8]. KGs help organize complex policy relationships, as in Poligraph, which structures privacy statements and improves entity recognition [6]. Our approach builds on our prior

work [2, 11] by extending structured representation and employing Sentence-BERT for semantic normalization [12], enabling meaningful cross-policy comparisons.

Large language models (PolicyGPT) have demonstrated success in zero-shot classification of policy sentences [14], although they lack graph-based interpretability. Unsupervised topic models such as LDA and BERTopic [4, 7] reveal hidden themes but compromise regulatory alignment and actor specificity, and lack cross-company and cross-regulatory comparisons. Most studies analyze policies in isolation, making it difficult to assess compliance across frameworks like GDPR, CCPA, and HIPAA [6, 13]. Persistent challenges include vague language, inconsistent labeling, and limited user-centric presentation [1, 5, 15]. Comparison of legal frameworks is rare, leaving regulatory alignment and jurisdictional differences unclear. Our approach identifies systemic ambiguities and non-compliance trends, and assesses the strength of privacy policies across various sectors.

3 Methodology

The GraphDPR Analysis Framework addresses the research questions by: (i) constructing semantically normalized KGs across organizations; (ii) applying regulation-aware embeddings and LDA topic modeling to the paths within these graphs; and (iii) enabling interpretable comparisons of how different organizations disclose and structure their data practices.

Poligraph framework focused on a single policy, we extend it by employing advanced sentence-level analysis so that the KG captures consolidated multiple domain-specific companies' PPs. The *.graphml file produced by Poligraph tool, encodes nodes and edges to represent policy actors (e.g., "Walmart," "advertiser") and data entities (e.g., "email address," "iris data"), and relationships (COLLECT, SUBSUM, and SHARE). Each relationship is further annotated with policy snippets, providing context on the nature of data collection or sharing. The five major stages of GraphDPR framework address RQ1–RQ3 via structured graph construction, normalization, and topic modeling (Algorithm 1).

3.1 Graph-Based Extraction and Semantic Processing Engine

The PP URLs for Walmart, Temu, Instacart, Boxed, Target, Costco, and Publix are provided to the PoliGraph-ER [6] framework, which utilizes named entity recognition and dependency parsing to extract relationships from policy text, outputting structured GraphYML and GraphML files. The structured pattern extraction from each GraphYML uses graph traversal and semantic filtering. Each PP graph is a directed heterogeneous graph, where nodes represent unique concepts such as entities and categories. Data types and edges symbolize labeled relationships from the PP text. Our methodology highlights the recursive expansion of edges in SUBSUM relationships, illustrating taxonomic containment (e.g., email address as personal information). It uncovers hierarchical context, enabling structured triples to convey detailed insights and facilitate semantic abstraction in tasks such as topic modeling. All extracted paths are normalized into a structured tabular dataset, where each row contains a unique quadruple: **(Company) → (Entity) → (Category) → (DataPoint)**. To identify

Algorithm 1: Privacy Policy Topic Modeling Workflow

Input: Set of privacy policy URLs
Output: High-level data collection topics across companies
Step 1: Policy Graph Generation (RQ1, RQ2)
url_list ← Set of privacy policy URLs; *graph_yml* ← `PoliGraph-ER(url_list)`
foreach *policy in url_list* **do**
    parse text, extract entities, categories, data points, output `GraphYML`
Step 2: Structured Triple Extraction (RQ1, RQ2)
triplets ← Extract (Entity, Category, DataPoint) paths from GraphYML;
Apply hierarchical path resolution via `SUBSUM` edges;
foreach *company* **do**
    build (Company, Entity, Category, DataPoint) quadruples;
Step 3: Semantic Normalization (RQ1
embeddings ← `SentenceBERT(node_labels)`; Clusters using cosine similarity;
Normalize Entity, Category, and DataPoint labels across companies;
Step 4: Knowledge Graph Construction (RQ1, RQ2)
graph ← Initialize Neo4j graph;
foreach *normalized quadruple* **do**
    Add (Company) → [:USES] → (Entity);
    Add (Entity) → [:BELONGS_TO] → (Category);
    Add (Category) → [:COLLECTS] → (DataPoint);
Step 5: Topic Modeling and Interpretation (RQ1, RQ3)
text_corpus ← Linearize graph paths into sequences;
lda_model ← Apply LDA on *text_corpus* with $k = 20$ topics, *topics* ← Extract top terms, Interpret and map topics to shared vs. company-specific practices;

shared data collection patterns among e-commerce companies (RQ1), we align data practices across privacy policies. The company-specific relationships extracted as quadruples may vary in phrasing across documents. Companies often use diverse terminology to mean similar actions (e.g., email addresses or location data). Thus, we employ a text processing pipeline to standardize and compare the patterns presented in the architecture in Fig. 1.

We extracted fragments indicating data collection intentions (e.g., "we collect your payment details") and used them to derive entity and category names for a downstream sentence transformer model. We used the pretrained **(all-mpnet-base-v2)** model for semantic embedding and normalization of entity labels, to ensure lexical normalization and unify similar terms, which uses Sentence-BERT embeddings [12] to measure contextual closeness via cosine similarity in semantic space producing generalized entities and categories list, integrating a multi-company KG infrastructure.

3.2 Construction of Unified Knowledge Graph in Neo4j

We construct a unified KG (Fig. 2) by integrating the semantically normalized data collection triplets derived in earlier phases implemented using the *Neo4j* graph database platform, and model data relationships across companies in a schema-aware, heterogeneous format. The KG encodes both shared practices and organization-specific patterns in a graph structure that is semantically rich,

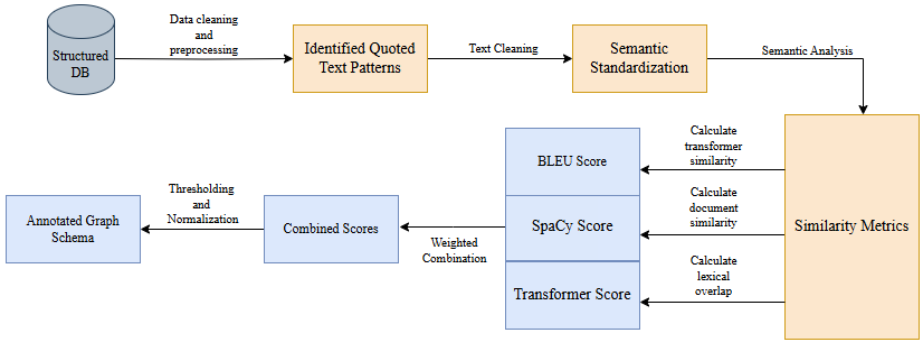

Fig. 1: Semantic Similarity Pipeline for Aligning Quoted Patterns in Privacy Policies

queryable, and visually explorable. Each node and relationship in the graph reflects its semantic role in the privacy data collection hierarchy. The core schema is shown in (Alg. 1, Lines 16-18).

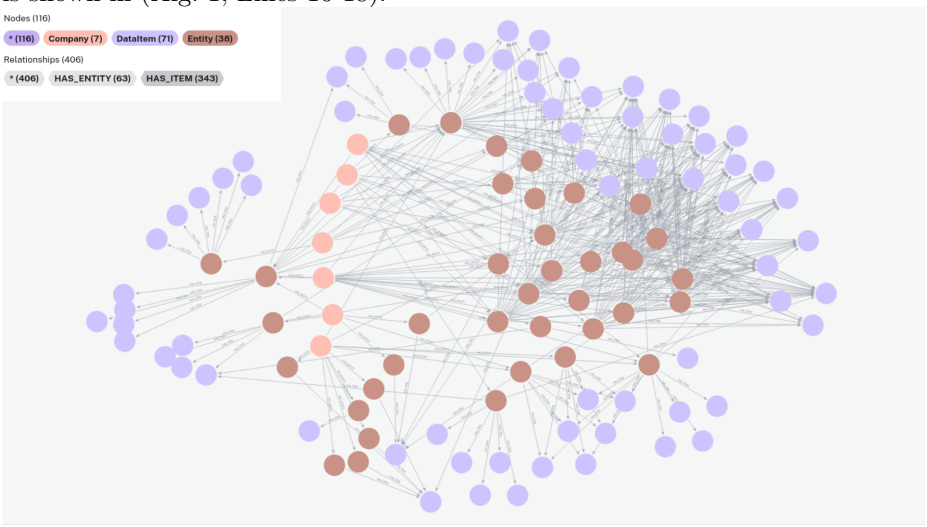

Fig. 2: Neo4j Knowledge Graph using Company-Entity-Category-Datapoint path

We constructed the KG using Cypher, Neo4j's declarative graph query language. Each normalized quadruple is ingested via Neo4j's Python driver, with constraints ensuring global uniqueness for shared nodes (e.g., `Category`, `DataPoint`), allowing company-specific instantiation of `Entity` nodes resulted in a densely connected, semantically enriched graph, where companies may share intermediate nodes and edges based on common data collection practices.

Multiple companies may point to the same `Category` node (e.g., "Personal Information") and the same `DataPoint` (e.g., "Email Address"), which reveals structural convergence in their policies. In contrast, unique edges such as those linking a company to uncommon (divergence) data types (e.g., "Retina Scan", "biometric disclosures") allow for the detection of privacy outliers (RQ1). Neo4j's *Cytoscape*-based extensions facilitate real-time graph exploration for investigative purposes. Its graph structure is adaptable, allowing for the integration of ad-

ditional semantic layers, including data-sharing partner information, legal data collection bases (for GDPR justification), or sensitivity levels. This flexibility provides a scalable foundation for privacy auditing and regulatory analysis.

3.3 Latent Topic Extraction Via LDA

For each company, data collection instances at the category level are extracted with contributing entities recorded to preserve attribution, yielding a normalized matrix that captures how each company engages with distinct categories of user data and identifies the actors involved in that collection. Target focuses on anonymized information, geolocation, and personal identifiers such as name, email, and purchase history. In contrast, Walmart exhibits a broader scope, collecting sensitive personal and biometric data—including face geometry, fingerprints, iris scans, and voiceprints—linked to entities like retailers and publishers. This contrast highlights Target's marketing-centric posture versus Walmart's deeper engagement with biometric and device-level tracking.

We apply LDA to discover latent thematic clusters from the aggregated data collection records. Each company's category-data point entries are linearized into a document to represent its PP behavior. These textual representations are vectorized using a Bag-of-Words model via `CountVectorizer` with standard English stopword removal. LDA model is trained using `LatentDirichletAllocation`. Two key results are: (i) topic-term distribution matrix for interpreting each topic based on top contributing keywords, and (ii) document-topic distribution matrix mapping each company to its associated topic mixture. These topic profiles enable us to cluster companies based on similarities in their data collection themes, such as *Contact and Identity Information*, *Device Metadata*, *Biometric Tracking*, and *Financial Transactions*. These topics can be used for comparative PP analysis, cross-industry benchmarking, and risk categorization.

4 Results and Discussion

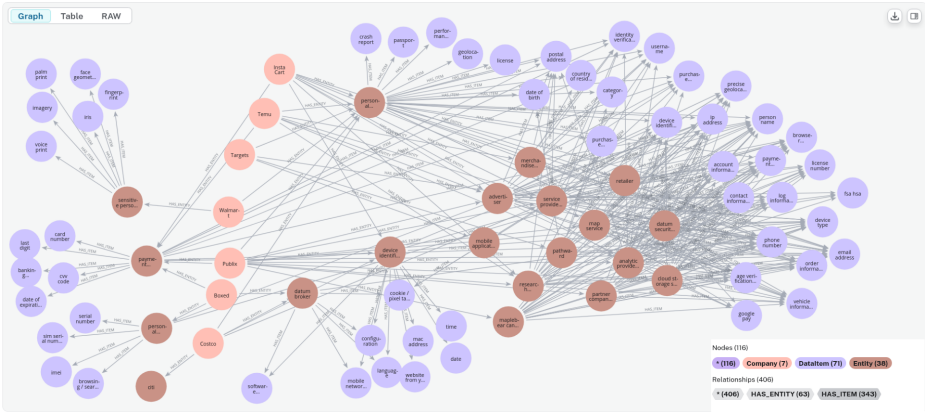

Fig. 3: Unified Multi-Company Privacy Policy Knowledge Graph

The knowledge graph of the integrated, cross-company graph combines Walmart, Publix, Costco, Boxed, Target, Temu, and Instacart PPs, providing insights into the structural, semantic, and behavioral aspects of data collection and sharing practices (RQ1). The graph reveals structured sharing edges to nodes

such as `advertiser`, `service provider`, and `Stripe`, confirming the external transmission of personal data as disclosed in the policy. This disclosure shows variable specificity in sharing, supporting discrepancy analysis (RQ2).

The macro-level insights from the unified merged graph (Fig. 3) are:Nodes like `device identifier`, `email address`, and `payment information` in most company subgraphs, highlight standardized collection practices. Related data elements (e.g., `IMEI`, `serial number`, `sim serial number`) grouped under abstract categories (e.g., `device identifier`) support semantic normalization across policy clusters. Biometric subtypes such as `iris` or `face geometry` included in some policies signal potential compliance inconsistencies.

Topic Modeling via LDA: Latent thematic patterns analyzed using LAD impacts the granularity of the extracted topics with $k = 5$ topics (manually interpreted based on top-ranking keywords per topic). Walmart shows a strong emphasis on biometric and financial data, reflecting its use of checkout technologies. Temu and Target prioritize transaction-related topics, suggesting deep integration with shopping behavior analytics. Costco centers on device metadata for tracking and fraud prevention, while Boxed emphasizes contact and transaction data typical of profiling-heavy platforms. Notably, InstaCart and Publix share identical topic profiles, indicating policy convergence. These topic patterns offer a deeper understanding of each company's privacy posture and facilitate a semantic comparison of PP strategies across e-commerce and retail sectors (RQ3).

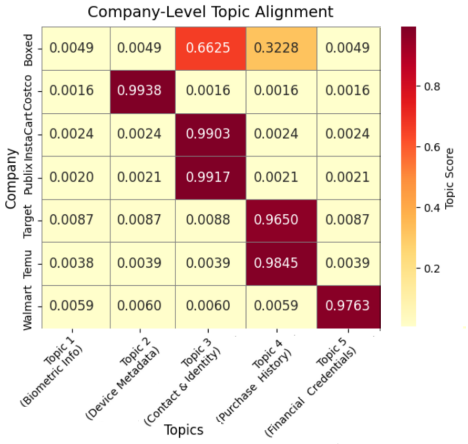

Fig. 4: Company-Level LDA Scores

5 Conclusion and Future Work

GraphDPR is a unified, graph-based framework for analyzing e-commerce privacy policies through semantic normalization, structural alignment, and topic-driven profiling. It extends POLIGRAPH to support multi-policy analysis and utilizes sentence transformer embeddings with Neo4j to combine diverse policy vocabularies into a coherent knowledge graph. LDA reveals five core data collection themes—biometric, device metadata, contact information, purchase history, and financial credentials, enabling the creation of interpretable, probabilistic company profiles. This combined structural and semantic analysis highlights shared norms (e.g., contact and tracking data) and company-specific behaviors (e.g., biometric and financial disclosures), while also identifying outliers that pose compliance risks. Future work includes embedding GDPR, CCPA, and HIPAA clauses for rule-based auditing and expanding to additional clause types, such as retention and legal justification, making GraphDPR a scalable, legally grounded solution for automated PP benchmarking.

References

[1] Adhikari, A., Das, S., and Dewri, R. (2022). Privacy policy analysis with sentence classification. In *2022 19th Annual International Conference on Privacy, Security and Trust (PST)*.

[2] Barrett, S., Ramasamy, V., Dorai, G., and Boswell, B. (2025). Navigating privacy policies with nlp and graph mining: Advancements in user-centric legal document analysis. In *SoutheastCon 2025*, pages 1013–1022.

[3] Binns, R., Lyngs, U., Van Kleek, M., Zhao, J., Libert, T., and Shadbolt, N. (2018). Third party tracking in the mobile ecosystem. In *Proceedings of the 10th ACM Conference on Web Science*, pages 23–31.

[4] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

[5] Costante, E., Sun, Y., Petković, M., and den Hartog, J. (2012). A machine learning solution to assess privacy policy completeness. In *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society*, WPES '12, page 91–96, New York, NY, USA. ACM.

[6] Cui, H., Trimananda, R., Markopoulou, A., and Jordan, S. (2023). Poligraph: automated privacy policy analysis using knowledge graphs. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA. USENIX Association.

[7] Grootendorst, M. (2022). Bertopic: Neural topic modeling with class-based tf-idf. In *arXiv preprint arXiv:2203.05794*.

[8] Keymanesh, M., Elsner, M., and Parthasarathy, S. (2020). Toward domain-guided controllable summarization of privacy policies. In *Natural Legal Language Processing (NLLP) Workshop at KDD 2020*.

[9] McDonald, A. M. and Cranor, L. F. (2008). The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Info. Society*, 4(3):543–568.

[10] Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33.

[11] Ramasamy, V., Barrett, S., Dorai, G., and Zumbach, J. (2025). Unveiling Privacy Policy Complexity: An Exploratory Study Using Graph Mining, Machine Learning, and NLP. ArXiv. Preprint, accepted AIRC 2025.

[12] Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.

[13] Shvartzshnaider, Y., Apthorpe, N., Feamster, N., and Nissenbaum, H. (2019). Going against the (appropriate) flow: A contextual integrity approach to privacy policy analysis. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 162–170.

[14] Tang, C., Liu, Z., Ma, C., Wu, Z., Li, Y., Liu, W., Zhu, D., Li, Q., Li, X., Liu, T., and Fan, L. (2023). PolicyGPT: Automated Analysis of Privacy Policies with Large Language Models. arXiv preprint.

[15] van der Schyff, K., Prior, S., and Renaud, K. (2024). Privacy policy analysis: A scoping review and research agenda. *Elsevier: Computers and Security*, 104065.