Business-government relations in the digital age

Corporate responses to policymaking in the European Union

Ocelík, V.

Publication date
2024

CHAPTER 3

Shifting Battlegrounds in Corporate Political Activity: The Dynamics of Information Strategies in the EU General Data Protection Regulation[4]

3.1 Introduction

Companies are increasingly lobbying the European Union (EU) in an effort to shape (forthcoming) policies (Coen & Vannoni, 2020; Hanegraaff & Poletti, 2021). This trend is particularly evident in the digital technology sector, which has emerged as the dominant EU lobby arena based on annual corporate expenditures (Bank et al., 2021). This surge can be attributed to the European Commission's (EC) proactive regulatory approach. The EC has declared its intention to "[…] manage the transformation into the digital age" (Von der Leyen, 2019: 14). Following the passage of the General Data Protection Regulation (GDPR) in 2016, legislative measures have expanded to include areas like AI, cyber security, high-performance computing, online platform regulation through the Digital Services Act (DSA), and regulations for 'gatekeeper' companies under the Digital Markets Act (DMA).

Given the extraterritorial reach of EU regulations, which is sometimes labelled the 'Brussels effect' (Bradford, 2020), what happens in the EU profoundly affects business operations both within and beyond the European Single Market (Coche et al., 2024). For example, the GDPR and its predecessor have spurred a wave of privacy rules in other countries and regions around the world. Such EU regulatory activities compel firms to engage in corporate political activity (CPA). This involves strategically navigating the 'nonmarket' environment (Baron, 1995a), and influencing governmental policies and decisions (Getz, 1997; Hillman & Hitt, 1999). CPA scholarship differentiates between three approaches to political strategy: financial, constituency-building and informational. The main difference between the approaches concerns the manner in which firms target policy makers, namely through supplying political candidates and parties with material resources (financial), through grassroots mobilization (constituency-

[4] An abridged version of this chapter has been published as: Ocelík, V., Kolk, A., & Irion, K. (2023). Shifting Battlegrounds: Corporate Political Activity in the General Data Protection Regulation. *Academy of Management Proceedings*, *1*, 11059.

building) or through supplying policymakers with intelligence.

Although CPA has been the subject of extensive investigation (Lenway et al., 2022; Mantere et al., 2009; Schuler et al., 2019), we argue that the current body of literature presents several significant gaps. To start with, CPA research has paid relatively limited attention to the *substance* of information strategies. There is a disproportionate focus on the outcomes of information strategies, such as the economic implications of filing anti-dumping petitions or providing congressional testimony (Marsh, 1998; Ridge, Ingram, Abdurakhmonov, & Hasija, 2019; Schuler, 1996), while overlooking the core content of these strategies themselves. This oversight might stem from the sparse representation of empirical EU studies in CPA literature. A recent review of major business journals spanning six decades (Lenway et al., 2022) underscored this, revealing only two EU-centric articles (Coen & Vannoni, 2020; Frankel & Højbjerg, 2012) – a stark contrast to the 28 U.S.-focused pieces. Interestingly, political science research indicates that effective EU lobbying hinges primarily on providing reliable, actionable intelligence to policymakers (Broscheid & Coen, 2007; Coen et al., 2021; Coen & Vannoni, 2020). This underlines the value of interdisciplinary collaboration for CPA research, a sentiment echoed by Lenway et al. (2022). Additionally, the CPA discourse, while centered on firms, has rarely considered the lobbying tactics of other stakeholder groups with opposing policy objectives. Entities like non-governmental organization (NGOs), public authorities, and trade associations lobby government officials as well, yet often have interests and ideal policies that diverge from those of corporate entities. As political scientists have pointed out, firms often 'compete' with these stakeholders in their lobbying (Bernhagen & Bräuninger, 2005), which implies that their role should be considered in a comprehensive analysis.

To bridge these identified gaps in contemporary CPA literature, our study poses the following central research question: How do firms adapt their information strategy in response to institutional procedures and information provided by other lobbyists? We have chosen the EU, with its GDPR as the focal point, as our empirical setting. The GDPR, being the precursor to a series of innovative digital regulation, provides a rich context for exploration. Delving into lobbying intricacies of the GDPR offers vital insights relevant to subsequent regulations like the DMA and the Digital Services Act (DSA). As previously emphasized, the EU presents an apt context for shedding light on

information strategies, an area that has been understudied and whose implications extend far beyond Europe's territorial confines.

Our data comprises 266 documents from 204 organizations, submitted to the EC during two GDPR consultation rounds. These documents are analyzed using structural topic modeling (Roberts, Stewart, & Tingley, 2014) – a topic modeling algorithm (Hannigan et al., 2019) that incorporates document metadata. This methodological approach aligns with Lenway et al.'s (2022, p. 31) broader observation that "the increasing ease with which methodologies such as natural language processing can be employed will create ample opportunity and copious data to explore the substance of information-based CPAs". The scalability and transparency of topic modeling algorithms allows us to efficiently analyze a formidable amount of textual data in a transparent and reproducible manner.

Our study makes three contributions to the extant CPA literature. First, we identify two distinct information strategies used by firms during the lobbying process: categorical information strategies and conditional information strategies, which are distinct in terms of their informational substance. During the initial phase, firms advocate for the comprehensive addition or omission of prospective regulations (categorical information strategy). Conversely, in the subsequent stage, lobbying efforts pivot to more nuanced aspects, as firms attempt to introduce caveats or exemptions to general rules (conditional information strategy). Consequently, we term these two policy stages as 'shifting battlegrounds'. Within this dynamic, firms first seek to influence what is included and excluded in the legislation, after which they need to engage the interests of other stakeholders.

Second, our analysis challenges prior findings (McKay & Yackee, 2007), by indicating that firms indeed scrutinize the lobbying activity of other stakeholders and adapt their informational outputs to policymakers in later policy stages accordingly. This suggests that interest groups have come to appreciate that they may advance their policy objectives by observing and responding to the demands of rival stakeholder groups during lobbying.

Third, drawing from the literature on interest group politics (Crombez, 2002; Eising, 2007), we elucidate the interplay between institutional procedures and information strategies. This focus responds to Lenway et al.'s (2022) appeal for further integration of political science insights to rectify prominent oversights in CPA research. As such, our study reveals the

importance of broadening the conceptualization of institutions within nonmarket strategy scholarship by underscoring the critical role of institutional procedures. CPA research has been profoundly shaped by the foundational work of Baron (1995), which predominantly casts institutions as governmental agencies, such as Congress and the Federal Trade Commission. Yet, our findings point out that businesses must navigate a wider spectrum of institutional constraints. Specifically, throughout the policymaking process, corporations encounter enhanced prospects to mold public policy prior to regulatory entities crystallizing and broadcasting their legislative aspirations. These institutional procedures are critical for firms to recognize and incorporate in their nonmarket strategies.

This chapter proceeds as follows. We commence by discussing pertinent CPA literature on firms' political strategies to sway policymakers (Hillman & Hitt, 1999). Here, we emphasize information strategies and the significance of institutional backdrops, bolstered by insights from political science research regarding the influence of the EU's legislative process on the lobbying dynamics of interest groups (Crombez, 2002; Eising, 2007). Subsequently, we offer an in-depth account of our empirical setting, data sources and analytical methodology. This leads into the presentation of our topic model analysis results, accompanied by a comprehensive explanation of our algorithmic interpretation. Concluding the chapter, we elaborate on the notion of 'shifting battlegrounds', wherein we delineate two distinct information strategies. We also identify potential constraints of our study, including the ethical implications of our methodology, and propose avenues for future research.

3.2 Theoretical Background

Scholarly work on CPA has significantly enriched our comprehension of the motives behind (Baysinger, 1984; Mitnick, 1981; Schuler, 1996), the methods of (Aplin & Hegarty, 1980; Boddewyn & Brewer, 1994; Frankel & Højbjerg, 2012; Hillman & Hitt, 1999; Meznar & Nigh, 1995; Schuler et al., 2002), the timing (Lenway & Rehbein, 1991; Rehbein & Schuler, 1999; Schuler, 1996), the locales (Sutton et al., 2021), and the impacts (Hadani et al., 2017; Hadani & Schuler, 2013; Hillman et al., 1999; Lux et al., 2011; Rudy & Cavich, 2020; Werner, 2017) of corporate actions to shape governmental policy. Although business lobbying has been studied by scholars from other disciplines, CPA

research distinctively probes these dimensions through the lens of corporate and managerial perspectives (Shaffer, 1995).

Guided by exhaustive reviews that chart the CPA discipline and its diverse facets (Lawton et al., 2013a; Mantere et al., 2009; Mellahi et al., 2016; Rajwani & Liedong, 2015; Sun et al., 2021), our examination homes in on two salient threads. The first encompasses literature considering corporate political strategies, specifically addressing the 'how' question in the CPA literature (cf. Getz, 1997). This segment especially emphasizes information strategies, given their relevance within the EU sphere (Broscheid & Coen, 2007; Coen et al., 2021; Coen & Vannoni, 2020). The second thread considers the role of institutions in CPA (Lim, 2022; Marquis & Qian, 2014; Oberman, 1993), broadening the scope to integrate insights from political scientists, particularly those illuminating the EU legislative process and its bearings on information strategies. We will discuss them consecutively below.

3.2.1 Corporate Political Strategies

The domain of corporate political strategies delineates the methods by which corporations attempt to influence governmental policy (Aplin & Hegarty, 1980; Boddewyn & Brewer, 1994; Bonardi & Keim, 2005; Coen & Vannoni, 2020; Holburn & Vanden Bergh, 2008; Meznar & Nigh, 1995; Nyberg & Murray, 2020; Ozer & Alakent, 2013; Rudy & Johnson, 2019). An instrumental framework, which has been widely adopted, adapted, and critiqued (e.g., Funk & Hirschman, 2017; Hadani et al., 2015; Lawton et al., 2013a; Lux et al., 2011; Mbalyohere et al., 2017), and that encapsulates these strategies was introduced by Hillman and Hitt (1999). Their framework proposes a triad of discernment. Companies must first ascertain the depth and nature of their political engagement. This can either be relational, characterized by sustained interactions and a broad engagement scope, or transactional, which tends to be temporary and pertains to specific issues or events. Subsequently, firms must decide on the modality of their political interactions. This can involve direct, individual engagement or a collective approach, for example through trade associations. Lastly, companies must determine their strategic approach. These strategies can be segregated into the following: financial strategies, encompassing campaign donations, contributions to political figures, or bestowing other materials perks;

constituency-building strategies, designed to leverage public opinion and include grassroots campaigns involving employees, and managing public relations; and, finally, information strategies, including traditional lobbying, generating research or technical documentation, and offering expert testimonies.

As explained above, this chapter focuses on information strategies, which firms employ in their attempts to influence public policy by providing policymakers information about policy positions and the costs and benefits of policy outcomes (Hillman & Hitt, 1999; Rajwani & Liedong, 2015). Policymakers often face a gap in intricate technical knowledge or struggle with assessing the economic or practical implications of their legislative proposals. Hence, they frequently rely on information from business to understand the potential ramifications of proposed legislation, especially within the convoluted realm of emerging digital technologies. In this landscape, corporations can step in as valuable repositories of domain-specific knowledge, and, ideally, can assist policymakers in avoiding pitfalls and unintended negative ramifications. Firms that consistently provide accurate information can establish themselves as reliable partners in the policymaking process, which not only grants firms a foothold in the political arena but also ensures their continued access and influence over time (Schuler et al., 2002).

Information strategies remain the least investigated corporate political strategy (Rajwani & Liedong, 2015). The relatively few studies on the role of information strategies have mostly focused on their antecedents and implications for firm performance, usually considering tactics such as filing anti-dumping petitions and providing congressional testimony in the U.S. context. Results showed filing petitions to be a function of domestic circumstances rather than foreign competition (Schuler, 1996), while the performance implications of petition filing appeared positive, provided that the petition was successful in the final stage (Marsh, 1998). Congressional testimony also seemed to yield abnormal financial returns, as they indicate to investors that firms boast influence in the political arena (Ridge et al., 2019). From the perspective of multinational enterprises, the landscape of information strategies takes on added complexities. Information strategies tend to be more often adopted in pluralist countries (Hillman & Wan, 2005) and when foreign subsidiaries rely on local intangible resources (Shirodkar et al., 2022), yet do not appear to increase the legitimacy of foreign subsidiaries

(Banerjee & Venaik, 2018). Furthermore, the rise of digital technologies has prompted scholars to theorize about their role in CPA. For example, Liedong et al. (2020) discussed the role of 'big data' – i.e., data that is large in volume, velocity, and variety – in corporate political strategies, and presented numerous propositions on how BDA increase CPA success.

Scholars have also started to clarify how firms use information strategies to shape the public discourse, which indirectly influences governmental policy. This research often adopts a critical perspective on the role of firms in politics (Néron, 2016), and seeks to elucidate how CPA challenges democratic government. On the issue of climate change, researchers have probed into the dynamics of 'information warfare' between firms, NGOs and social movements (MacKay & Munro, 2012), and have dissected how firms manipulate public debate to build a common identity with citizens and harmonize corporate and public interests (Nyberg et al., 2013). Furthermore, a collection of articles detailing the political contest between the Australian government and the mining industry has shed light on how firms utilize discursive strategies to leverage media coverage and (de)legitimize public policies, potentially posing challenges to core democratic principles (Murray & Nyberg, 2021; Nyberg & Murray, 2020). Finally, Murray and Flyverboom (2021) consider how corporate advocacy is evolving in the digital age, coining the term 'datafied CPA'. This refers to emergent informational forms of CPA that harness the capabilities of digital technologies. The authors posit that the accessibility, credibility, and context of information can be strategically and algorithmically tailored, carrying significant political implications.

3.2.2. Institutional Constraints on Corporate Political Strategies

Institutional theory underscores the pivotal role of societal norms and expectations in guiding both organizational and individual actions (DiMaggio & Powell, 1983; Meyer & Rowan, 1977). As articulated by Oberman (1993, p. 215), institutions "define conditions and set limits for maintaining a stable system; they regulate social relations to maintain conformity with existing value patterns and consistency among these patterns themselves". In simpler terms, institutions establish a structure that calibrates the expectations of interest groups, enhancing predictability by specifying what constitutes

appropriate lobbying conduct. Within CPA research, the significance of institutions has been invoked in various ways (Getz, 1997). One body of work suggests that CPA serves as a conduit for organizations to secure institutional resources and bolster their legitimacy (Baysinger, 1984; Oberman, 1993; Shaffer, 1995), while another posits that organizations with CPA structures are likely to use CPA because they consider it an integral part of their strategic toolkit (Baysinger, 1984; Rehbein & Schuler, 1995, 1999). More recently, scholars have revisited the role of institutions while analyzing CPA in emerging markets and weak institutional environments (Dorobantu et al., 2017; Marquis & Raynard, 2015; Mbalyohere et al., 2017; Mbalyohere & Lawton, 2018; Rodgers et al., 2019). They broaden earlier discussions on how institutional variances across nations, such as form of government and the organizational form of stakeholder groups, induce disparities in business-government relations (Hillman & Keim, 1995).

The nexus between institutional theory and CPA lies in the proposition that firms' alternatives of political actions are circumscribed by the prevailing institutional frameworks in their operational milieu (Getz, 1997). An accumulating body of evidence underscores the tenet that institutions demarcate the boundaries of acceptable lobbying conduct by corporations, both in domestic and international settings (Barron, 2011; Dorobantu et al., 2017; Hillman, 2003; Hillman et al., 2004; Hillman & Wan, 2005; Kolk & Pinkse, 2007; Lawton et al., 2013a; Ozer & Alakent, 2013; Rugman & Verbeke, 1998; Schuler et al., 2002; Sun et al., 2021; Zhang et al., 2016). However, CPA scholars mostly conceptualize institutions as authoritative bodies entrusted with a mandate to craft and implement public policy, such as parliaments, courts, government agencies and international organizations. What remains relatively underexamined are the institutional mechanisms – namely, the formalized protocols governing legislative enactment (for a notable exception, see Somaya & McDaniel, 2012). It is in this gap that political science scholarship offers vital guidance, as scholars in this field, particularly those studying lobbying dynamics in the EU, have discerned that institutional procedures affect which policymakers are targeted by lobbyists.

Crombez (2002) constructed a comprehensive framework elucidating lobbying avenues within the EU. He concluded that during the initial proposal phase of policy formulation, lobbyists should target policymakers with preferences closest to their own. However, as the process progresses to the

voting phase, the focus should pivot towards the critical policymakers. Relatedly, Klüver (2011) suggested that, because the regulatory proposal of the EC acts as a foundational blueprint for subsequent deliberations between the Council and the EP, it becomes challenging for other stakeholders to introduce substantial alterations. Consequently, it is during the policy formulation stage that lobbyists can exert most influence. This sentiment finds consonance with Marshall (2010), who articulated that lobbying entities typically initiate their activities at the earliest juncture in the legislative trajectory in order to frame the debate.

Shifting the focus to the EP, a number of studies have expounded on the procedural intricacies influencing lobbying tactics. Varela (2009) contended that the consultation procedure prompts lobbyists to interface with the EP, with the aim of bolstering the likelihood that the EP will include their grievances in its opinion on the EC's proposal. Likewise, the expansion of the co-decision procedure (when the EP became co-legislator, with the Council, under the Treaty of Lisbon) precipitated a marked intensification in lobbying engagement, with lobbyists acutely cognizant of the EP's amplified significance in the legislative process (Huber & Shackleton, 2013). The EP's committee-driven architecture further nuances lobbying maneuvers. Lobbyists calibrate their outreach depending on the specific phase within each committee process (Marshall, 2010). Moreover, committees that exhibit a higher proportion of Ordinary Legislative Procedures (where the EP co-legislates on an equal footing with other EU institutions) to Own Initiative Reports (where the EP produces a resolution that addresses a specific issue) witness greater private interest mobilization (Coen & Katsaitis, 2019).

To sum up, political science furnishes us with two pivotal vantage points essential to understanding information strategies within the EU context, and on which we will build in our empirical exploration below. First, the extent and nature of lobbying endeavors are anticipated to oscillate based on the distinct stages of the legislative procedure. Second, once the EC publishes a draft proposal or an official policy communication, it becomes a reference point for ensuing dialogues between interest groups and EU institutions. The draft proposal demarcates the contours of all subsequent negotiations: it becomes an institutional anchor that circumscribes the ambit of demands interest groups can proffer during their advocacy efforts. Given this, persuading the EC to jettison specific stipulations or tenets enshrined in the

proposal wholesale becomes untenable. Hence, the substance of their information strategies needs to be dynamically readjusted to resonate within this institutional framework.

3.3 Context, Data and Methodology

This section commences with a brief overview of the EU legislative procedures, focusing on European regulation, with an emphasis on the GDPR. The term 'regulation' in this context specifically denotes laws that are instantly enforceable across all EU member states. This is distinct from 'directives' that necessitate translation into national law. Subsequently, we discuss our data and methodological approach employed in this study.

3.3.1 Research Context

Within the EU, the EC has the exclusive right to set the legislative agenda: it drafts legislation that then has to be negotiated and approved by other European institutions, specifically the Council and the EP (Wallace et al., 2015). Notably, prior to tabling new regulations, the EC solicits feedback from various stakeholders. At this juncture, there is an absence of publicly available information that outlines the substance, scale and scope of the proposed regulation. Upon integrating this feedback, the EC issues a communication to the other EU institutions in which it clarifies primary objectives of the impending legislation. Crucially, it is only here that interest groups obtain insights into the aspirations of other stakeholders, discerning what they aimed to incorporate or exclude from the upcoming legislation, and how the EC has parsed these preferences. Following the release of this communication, interest groups can again submit lobby documents to the EC, detailing their position on the outlined objectives, and attempting to sway the EC towards certain policies. Subsequent to this stage, the EC publishes its formal legislative proposal, setting the stage for further deliberation between the Council and the EP.

The GDPR, the primary subject of this study, underwent this intricate procedure. Characterized as the most heavily lobbied regulation in the annals of the EU, the GDPR legislative journey was both lengthy and intense. Starting from the publication of the legislative proposal, the entire process spanned over two years, entertained close to 4000 proposed amendments, and then

required an additional two years for its culmination and formal adoption (D'Cunha, 2020). Such an extended and involved process underscores its pivotal role in the nonmarket environment of businesses. The GDPR emerged as a refinement and extension of the 1995 Data Protection Directive, which was beleaguered by issues related to poor compliance and enforcement (Granger & Irion, 2018; Hoofnagle, Van der Sloot, & Zuiderveen Borgesius, 2019). In essence, the GDPR embodies a robust data governance framework that compels companies operating within the EU to treat privacy and data protection with the same gravitas as anti-trust laws and foreign corrupt practices regulations (Hoofnagle et al., 2019). Whereas previously organizations de facto enjoyed the freedom to cast a wide net in cyberspace and harvest all personal data they deemed remotely relevant for current or future purposes, the GDPR incentivizes them to take a more deliberate, responsible and parsimonious approach to data collection and analysis.

There are a few empirical articles on (business) lobbying in the context of the GDPR, and these have primarily been published in journals centered on public policy and European politics. Kalyanpur and Newman (2019) analyzed lobbying dynamics that characterized the GDPR's legislative process. They postulated that the issue salience of data protection ushered in legitimacy challenges for European institutions. This effectively undermined traditional sources of political power, such as financial capital, organizational infrastructure and informational expertise. As a result, attributes like corporate size, global footprint, and substantial financial reserves, which were once assets in lobbying, ironically became liabilities (cf. Dür & Mateo, 2014). Another notable investigation on the GDPR was conducted by Atikcan and Chalmers (2019), who addressed the question of the criteria that drive interest groups to choose their affiliations on specific policy matters. Their findings suggest that interest group preference alignment is largely influenced by industry-specific costs of regulatory alterations. Finally, Christou and Rashid (2021) centered their inquiry on the 'right to be forgotten'. Their study highlights the prominence of issue salience and institutional polarization in shaping interest group's lobbying success.

3.3.2 Data Collection

In our study, the documents used as input for the topic modeling algorithm

were accessed through a Freedom of Information (FOI) request submitted to the EC. Specifically, we sought all pertinent documentation – a broad spectrum encompassing e-mails, correspondence, notes, agendas, presentations, and other relevant textual data – that the Commission had received in relation to the legislative procedure of the GDPR. The time frame spanned from January 2009 through June 2011.

These documents consist of replies to two open consultations. The first is the consultation concerning the EC's 'Comprehensive Approach on Personal Data Protection in the European Union' (CAPDP), while the second pertains to the 'Legal Framework for the Fundamental Right to the Protection of Personal Data' (FRPDP). The structure of the CAPDP was based on a three-part questionnaire where the EC sought stakeholders' perspectives on three main challenges: identifying emerging challenges in personal data protection; assessing if the existing legal framework was adequately meeting those challenges; and pinpointing potential avenues for remedial action. Conversely, the FRPDP was more focused on delineating prospective policy trajectories, particularly reflecting the feedback garnered from the CAPDP consultation. However, both consultation rounds took place before the EC published its legislative proposal for the GDPR in January 2012. The CAPDP documents dated around the period of January 2010, while the FRPDP document were timestamped around January 2011. Together, these two consultations received a total of 491 documents, authored by companies, NGOs, (national) public authorities, or trade associations.

In preparation for out topic modeling, an essential step was curating the corpus from the initial pool of documents we had accessed, as not all were suitable for the algorithm. A portion of the documents was written in languages other than English, specifically German and French. Given the algorithm's reliance on term co-occurrence patterns, it is imperative to have a uniform language corpus. As a result, we removed a total of 35 documents. However, we must emphasize that this removal does not introduce significant bias in our data, as most often (except for six documents) the same documents were submitted in both German and English, or French and English. Furthermore, some documents were too short or devoid of content to include. After filtering for these issues, we were left with 266 documents, 156 from the FRPDP and 110 from the CAPDP.

We then proceeded to extract the paragraphs from the documents to

create our final two corpora. We categorized the different author types into one of four options: companies (e.g., Facebook), NGOs (e.g., European NGO Alliance for Child Safety Online), trade associations (e.g., American Chamber of Commerce), and public authorities (e.g., European Medical Agency). *Table 3.1* summarizes this information.

Table 3.1. Data sources by author category

Consultation Round	Time period	Number of paragraphs	Company	NGO	Trade association	Public authority
Comprehensive approach on personal data protection in the European Union	2009	2576	619	292	1083	582
The legal framework for the fundamental right to the protection of personal data	2010-2011	3924	1253	443	1743	494

3.3.3 Structural Topic Modeling

Topic modeling constitutes a form of unsupervised machine learning based on the Bayesian statistical technique of Latent Dirichlet Allocation (LDA) (Blei, 2012; Blei et al., 2003; Blei & Lafferty, 2007). Topic modeling is increasingly used by management scholars to reveal phenomenon-based constructs and grounded conceptual relationships in textual data (Hannigan et al., 2019). Although it remains a relatively 'obscure' form of text analysis in management, there is a growing number of published articles in top-tier business journals explaining topic modeling in great detail (Bao & Datta, 2014; Croidieu & Kim, 2018; DiMaggio et al., 2013; Giorgi & Weber, 2015; Haans, 2019; Hannigan et al., 2019; Huang et al., 2018; Kaplan & Vakili, 2015). We refer to these sources for the generic approach, to instead briefly clarify the added value of *structural* topic modeling, which has become more frequently adopted by management scholars in recent years (Innis, 2022; Karanović et al., 2021).

Following the seminal publication of Blei et al. (2003), other work has

refined and extended the original algorithm. For example, Blei and Lafferty (2007) introduced the *correlated* topic model, which allows users to estimate the correlation between topics within documents. Put simply, this helps researchers to answer the question of how often a particular topic, such as international transfers of personal data, co-occurs within documents with another topic, such as notification obligations to supervisory authorities. The key innovation of *structural* topic models is that it permits researchers to incorporate document metadata into the topic model and estimate relationships between this metadata and topical prevalence and topical content (Roberts, Stewart, & Tingley, 2014). Here, topical prevalence denotes the degree to which a document is associated with a topic, while topical content refers to the words used within topics. With metadata, we refer to information about the document itself: the author, publication date, etc. Basically, this allows researchers to estimate how often a particular topic, such as technological neutrality, occurs within a document written by, for example, a trade association. The incorporation of metadata into the topic model enables a more structured interpretation of the underlying themes in the corpus. For our analysis, we incorporated the type of author of each document, i.e., whether it was authored by a company, NGO, public authority, or trade association.

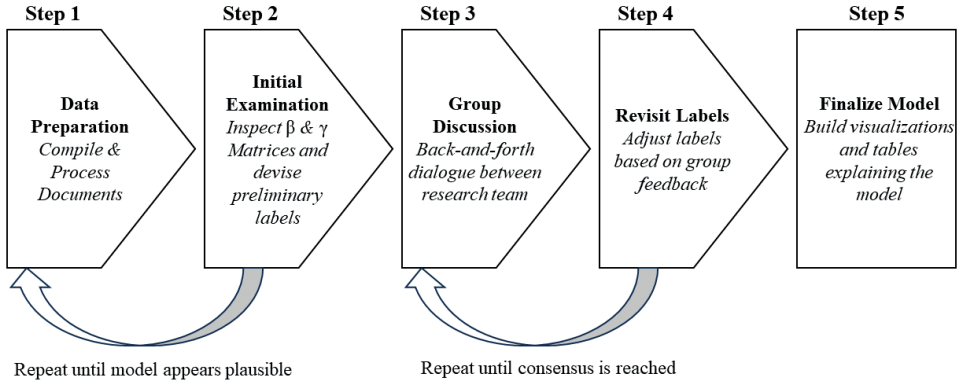
We performed our analysis in RStudio version 4.1.2. (R Core Team, 2021). The main software package we utilized is the *stm* package (Roberts, Stewart, & Tingley, 2014). We ran several models with different pre-processing parameters and number of topics. While we are well aware of established principles in text mining (Kobayashi et al., 2018), we also wanted to keep an open mind on what parameters might yield the most illuminating model of the documents (Schofield et al., 2017). We thus experimented with both lemmatization and stemming; term-document count threshold (the threshold at which terms are excluded from the topic model due to (in)frequency); and the inclusion of n-grams. We also experimented with (un)nesting the paragraphs into sentence and full documents. To assess what would be a reasonable number of topics to compute, we consulted commonly used performance metrics, such as the exclusivity and coherence of the topics (Roberts et al., 2016; Roberts, Stewart, & Tingley, 2014; Roberts, Stewart, Tingley, et al., 2014). We also visualized the topic models using the LDAvis visualization package intertopic distance map (Sievert & Shirley, 2014) to obtain a global overview of our models. These tools offered good estimates of

what number of topics would be useful for our purposes.

We proceeded to explore the two main matrices generated by the algorithm, namely the terms-over-documents matrix (β) and the topics-over-documents matrix (γ). The β-matrix contains the weighted term for each *k* topic, where the top weighted terms denote a latent structure. The γ-matrix contains the weighted documents for each *k* topics, and can thus be used to retrieve documents that load strongly onto a specific topic. The weights in both matrices denote probabilities, which logically sum to 1. The ultimate decision on what number of topics to select to represent a particular corpus should be made on how well the model elucidates underlying themes in the documents: a 'correct' number of topics does not exist (DiMaggio et al., 2013; Hannigan et al., 2019).

Topic modeling hinges on the systematic manipulation of data matrices. However, algorithmically identifying topics is just the initial step; the real expertise lies in interpreting and labeling them. As the algorithm provides clusters of co-occurring terms deemed as topics, it is the task of the research team to discern coherent narratives or themes within them. The task of labeling topics is an intricate back-and-forth of data interpretation and domain expertise. We delved into the most weighted terms for each topic, drawing upon the context provided by the terms-over-documents and topics-over-documents matrices. We assessed the significance, relevance and relationship of these terms, and, when necessary, revisited original documents for clarity. To ensure rigor and robustness, we engaged in back-and-forth discussions, challenging each other's interpretations, suggesting alternate labels and negotiating nuances. This iterative dialogue was paramount, as it ensures the final topic labels are both data-driven and contextually meaningful. Through these repeated cycles of analysis, debate and consensus-building, we transformed algorithmic outputs into insightful topic labels that resonate with the latent thematic landscape of the data corpus. We visualize this process in *Figure 3.1*. In the end, we settled on a model with 18 topics for the CAPDP corpus, and a model with 26 topics for the FRPDP corpus. We will now proceed to discuss the results of the structural topic model.

Figure 3.1. Step-by-step process of building a Topic Model



3.4 Results

3.4.1 Comprehensive Approach on Personal Data Protection in the European Union

Figure 3.2 plots the 18 topics on the vertical axis arranged by their overall prevalence in the entire corpus on the horizontal axis. The text next to each horizontal bar denotes the top 8 terms for that topic. We see, for example, that topic 9 (technology, challenge, …, society) accounts for 8 percent of all documents in the corpus. Similarly, topic 2 (share, public, …, exchange) accounts for approximately 4 percent of the entire corpus. Furthermore, the topics appear to be distributed rather uniformly across the corpus, with the vast majority of topics accounting for about 4-8 percent of all the documents in the corpus. Next, we filtered out the top 10 documents for each topic based on the topics-over-documents matrix. We then proceeded to label each topic, thus describing what each particular topic was about. As mentioned above, this involves a back-and-forth process where one moves from topics to documents. We were aided in this process by the expertise of legal scholars with intimate knowledge of the GDPR lobbying process. Any disagreements regarding the labels were discussed by the multidisciplinary co-author team until they were satisfactorily resolved.

As pointed out in the previous section, structural topic models allow for the estimation of relationships between arbitrary metadata and both the prevalence of topics and topical content (Roberts, Stewart, & Tingley, 2014). In order for us to understand the important themes for the different interest groups, we estimated the correlation between the type of author and the prevalence of each topic. We plot the topic prevalence for each author type in order to allow for author prevalence comparisons in *Figure 3.3*. Here, we notice that some topics occur in documents produced by all types of authors or groups of author types, while other documents appear rather exclusively in documents authored by a particular author. For example, all author types address the fragmentation of data protection regulation in their lobby documents, while data challenges for banks seem to be an important theme in lobby documents produced by trade associations. Finally, *Table 3.2* lists the 18 topics in descending order by their prevalence, along with their 8 most frequent terms; and the label we have assigned to that topic to denote its underlying theme.

Figure 3.2 18 topics from the "comprehensive approach on personal data protection in the European Union" corpus, arranged by prevalence

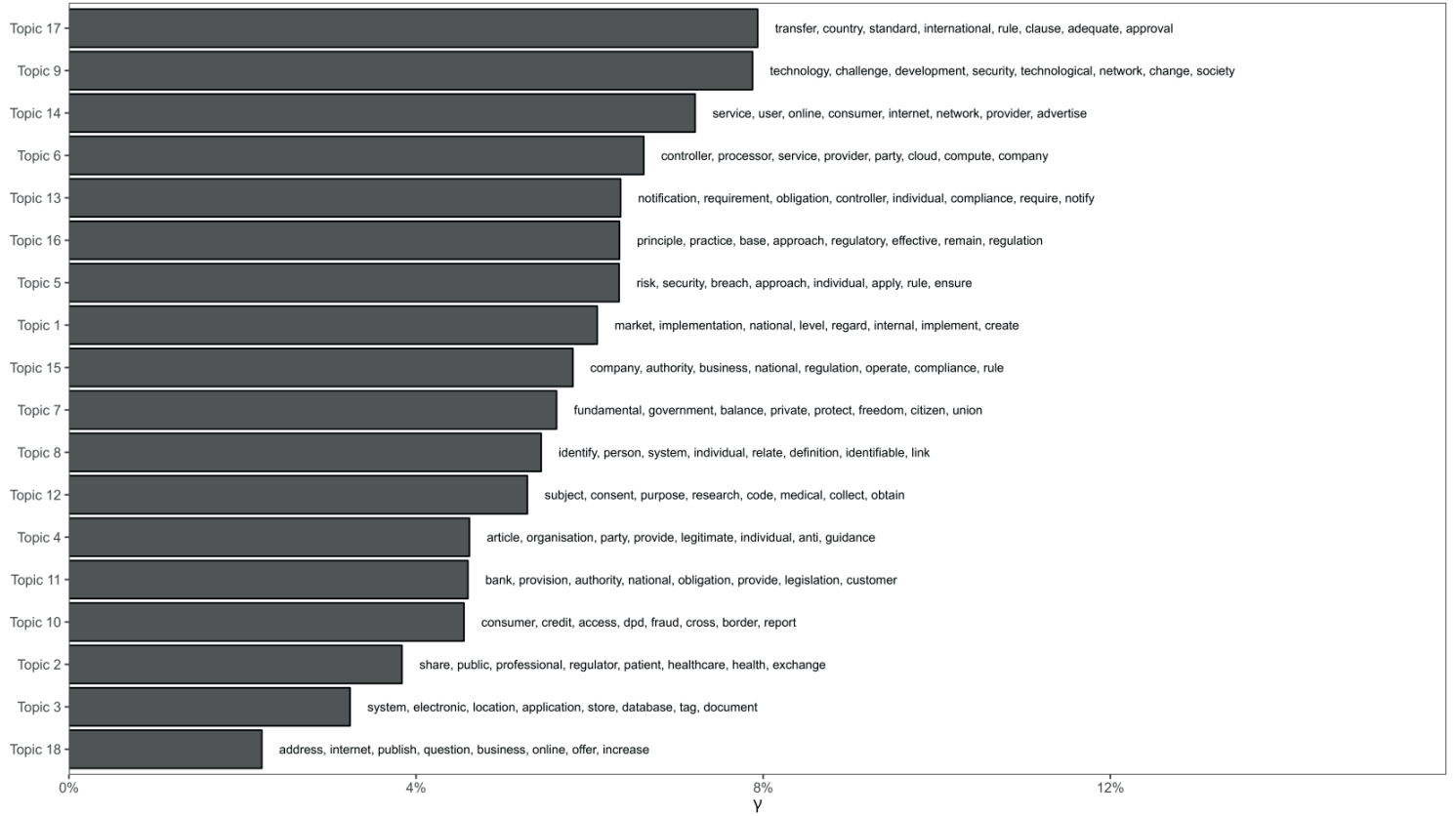


Figure 3.3. 18 topics from the "comprehensive approach on personal data protection in the European Union" corpus with estimated author prevalence

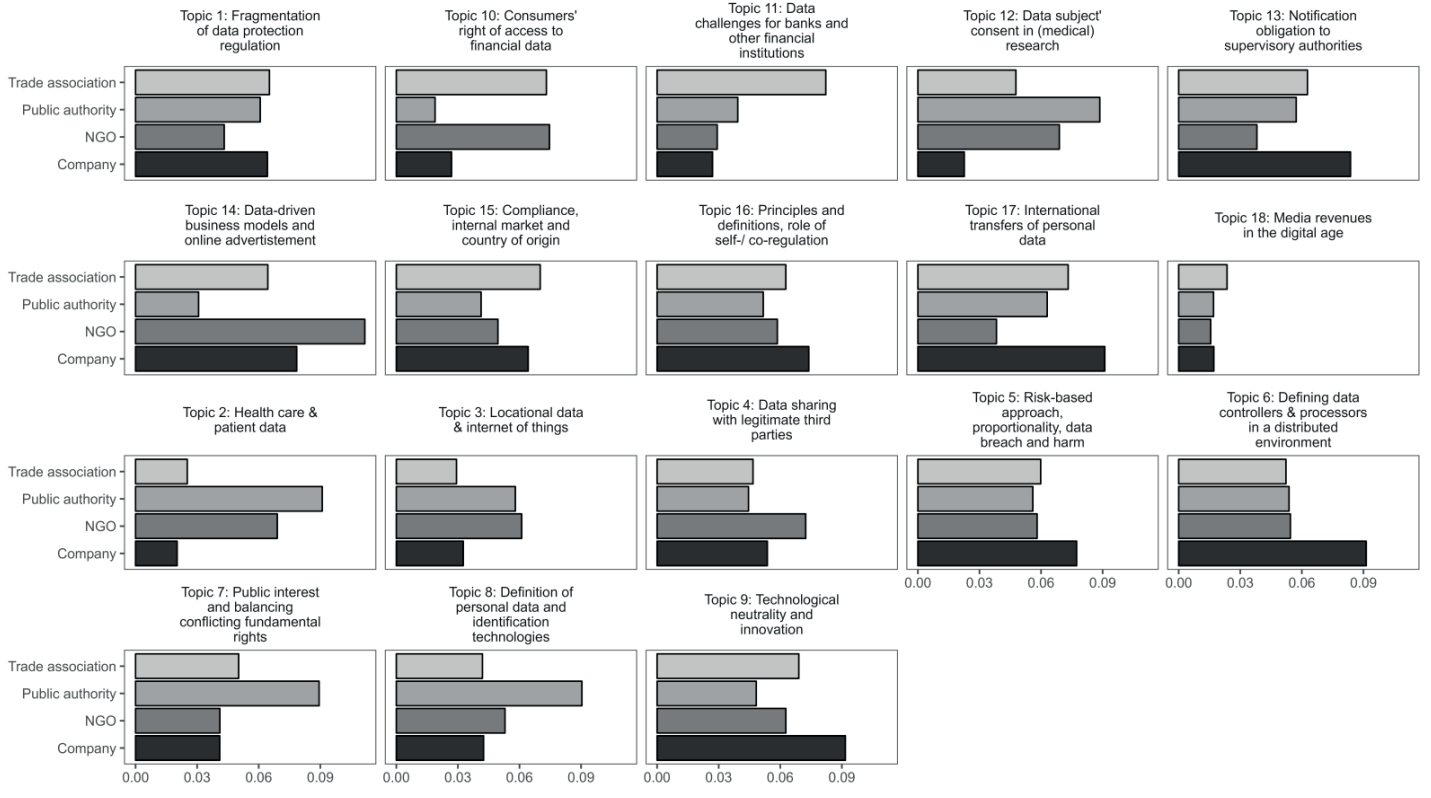


Table 3.2. 18 topics from the "comprehensive approach on personal data protection in the European Union" corpus, including their top terms and assigned label

Topic	Top terms	Assigned label
17	transfer, country, standard, international, rule, clause, adequate, approval	International transfers of personal data
9	technology, challenge, development, security, technological, network, change, society	Technological neutrality and innovation
14	service, user, online, consumer, internet, network, provider, advertise	Data-driven business models and online advertisement
6	controller, processor, service, provider, party, cloud, compute, company	Defining data controllers & processors in a distributed environment
13	notification, requirement, obligation, controller, individual, compliance, require, notify	Notification obligation to supervisory authorities
16	principle, practice, base, approach, regulatory, effective, remain, regulation	Principles and definitions, role of self-/ co-regulation
5	risk, security, breach, approach, individual, apply, rule, ensure	Risk-based approach, proportionality, data breach and harm
1	market, implementation, national, level, regard, internal, implement, create	Fragmentation of data protection regulation
15	company, authority, business, national, regulation, operate, compliance, rule	Compliance, internal market and country of origin
7	fundamental, government, balance, private, protect, freedom, citizen, union	Public interest and balancing conflicting fundamental rights
8	identify, person, system, individual, relate, definition, identifiable, link	Definition of personal data and identification technologies
12	subject, consent, purpose,	Data subject' consent in

	research, code, medical, collect, obtain	(medical) research
4	article, organisation, party, provide, legitimate, individual, anti, guidance	Data sharing with legitimate third parties
11	bank, provision, authority, national, obligation, provide, legislation, customer	Data challenges for banks and other financial institutions
10	consumer, credit, access, dpd, fraud, cross, border, report	Consumers' right of access to financial data
2	share, public, professional, regulator, patient, healthcare, health, exchange	Health care & patient data
3	system, electronic, location, application, store, database, tag, document	Locational data & internet of things
18	address, internet, publish, question, business, online, offer, increase	Media revenues in the digital age

3.4.2 Consultation on the Legal Framework for the Fundamental Right to Personal Data Protection

Figure 3.4 plots the 26 topics on the vertical axis and their overall contribution to the entire corpus on the horizontal axis. We see that topic 6 (national, authority, …, legislation) accounts for approximately 6 percent of the entire corpus. Similarly, topic 19 (risk, assessment, …, german) constitutes roughly 3 percent of the entire corpus. Interestingly, the topics appear less equally distributed in the FRPDP compared to the CAPDP corpus. The top 5 topics appear to account for a significant amount of the underlying themes in the corpus, while the bottom 3 appear to be niche subjects reserved for specific documents. Furthermore, the topics themselves appear already to be more coherent compared to the CAPDP corpus, as the top terms themselves are already quite indicative of a specific underlying theme. In a sense, we can already observe that the subject matter has become increasingly specific.

We then again paired each topic with the 10 documents that loaded most strongly onto that specific topic, and used both the topics top terms and documents to devise labels for each topic. Compared to the CAPDP corpus,

this was a much more straightforward process. Aside from the suggested top terms, the documents were also particularly elucidating of the underlying theme that the topic captured. As such, devising labels for the FRPDP corpus was less challenging than the CAPDP corpus. However, we still double-checked all labels carefully and ensured that there were no disagreements within the author team regarding their substance. Next, we again proceeded to estimate the relationship between author type and topic prevalence, and plotted the topic prevalence and author type (for the results, see *Figure 3.5*). Finally, we combined all 26 topics, including their top terms and labels, in *Table 3.3*.

Our analysis brought to light marked divergences in thematic substance across the consultation phases. During the initial consultation round, companies, NGOs, public authorities and trade associations lobbied for the inclusion or exclusion of policy proposals. However, during the second consultation, where there were a number of regulatory ideas articulated and consulted, the lobbying strategy shifted towards limiting or expanding the proposed set of regulatory ideas. The crux of this transformation lies in the fact that the crystallization of regulatory proposals and concepts geared towards updating data protection frameworks effectively functioned as an institutional anchor, shaping and confining the aspirations that stakeholder factions could present to the EC.

As such, there are two major differences between the two consultation rounds. First, as can be found in *Table 3.2* and *Table 3.3*, there were far fewer topics in the first consultant round than in the subsequent round. Second, as can be seen in *Table 3.4* and *Table 3.5*, there is a remarkable difference in the level of detail between the two consultation rounds. Whereas in the first round, we see abstract conversations about principles and the balancing of different fundamental rights, subsequent topics consider the finer nuances of data protection and privacy, such as how to categorize sensitive data, and which organizations must appoint a data protection officer. It is here that we see that discussions no longer center on the question whether such categories should exist in the first place, or whether data protection officers are necessary, but rather on what constitutes sensitive data, and where the boundary condition should be placed on mandating organizations to appoint a data protection officer.

Table 3.3. 26 topics from the "legal framework for the fundamental right to the protection of personal data" corpus, including their top terms and assigned label

Topic	Top terms	Assigned label
15	national, authority, level, harmonisation, implementation, ensure, enforcement, legislation	Harmonizing data protection regulation
20	technology, change, business, development, regulation, technological, effective, consumer	Technological neutrality to stay futureproof
22	principle, design, accountability, measure, concept, security, technology, support	Accountability and Privacy-by-design
6	notification, burden, system, administrative, reduce, requirement, form, authority	Administrative burdens
17	breach, notification, action, harm, sanction, requirement, sector, authority	Sanctions and data breach notifications
1	subject, applicable, request, access, apply, establish, provision, exercise	Clarify which rules are applicable
11	standard, notice, industry, transparency, support, form, international, global	Global privacy and industry standards
24	purpose, legitimate, forget, collect, principle, article, subject, minimisation	Purpose limitation and data minimization
16	transfer, country, adequacy, processor, bcrs, clause, bind, corporate	Simplifying the EU international data transfer regime
5	user, network, social, service, internet, control, online, content	Data subjects' control and informed consent
23	company, organisation, business, market, internal, officer, compliance, dpo	Organizations appointing a data protection officer
18	concept, responsibility, definition, identify, address, processor, responsible, approach	Defining personal data and clarifying responsibilities
12	consent, inform, context, opt, subject, require, explicit, obtain	The role of context in informed consent
13	regulatory, certification, initiative, industry, regulator, regulation, scheme, compliance	Promoting self-regulatory mechanisms
3	transfer, cloud, border, international, flow, compute, country, cross	Managing international data transfers and cloud computing
8	consumer, awareness, raise, activity, credit, access, bank, financial	Awareness raising across citizens and consumers
10	fundamental, article, freedom, balance, protect, court, human, expression	Balancing the fundamental rights of freedom, security, and justice

25	service, provider, customer, online, portability, application, company, product	Service providers and data portability
2	sensitive, category, context, type, definition, list, special, genetic	Categories of sensitive data
19	risk, assessment, impact, insurance, company, method, require, german	The importance of risk in data protection impact assessments
7	key, research, code, party, consultation, stakeholder, opinion, issue	Integrating key stakeholders' positions
21	health, public, patient, safety, care, medical, share, system	Health care & patient data
14	child, online, age, people, world, card, internet, site	Protecting children online
9	public, record, concern, citizen, private, religion, worker, employment	State records of religious beliefs
4	obligation, professional, independent, task, specific, subject, official, duty	The obligations and role of data protection officers
26	review, dpas, enforcement, sufficient, power, limit, include, resource	The powers and resources of data protection authorities

Figure 3.4. 26 topics from the "legal framework for the fundamental right to the protection of personal data" corpus arranged by prevalence.

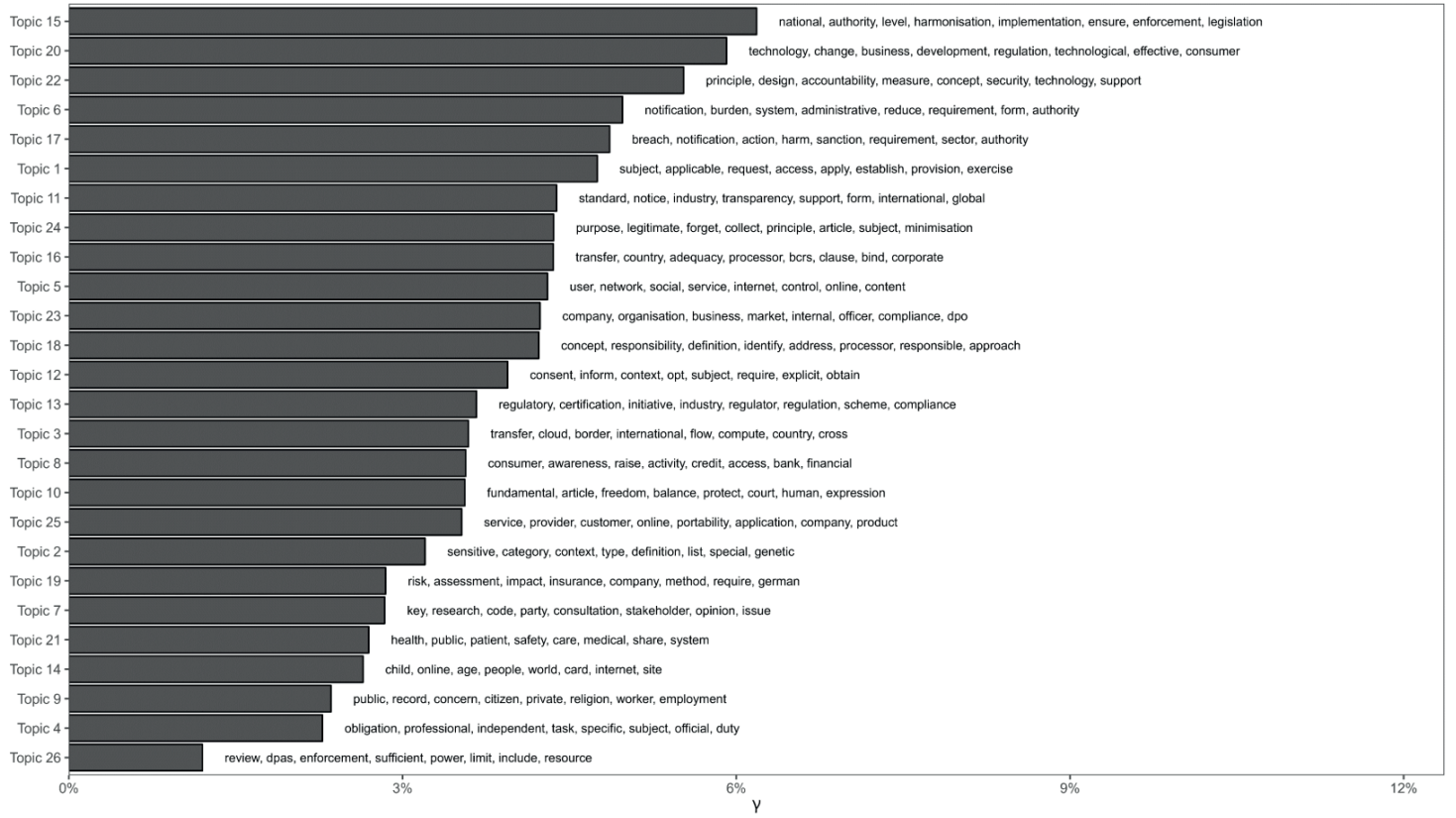
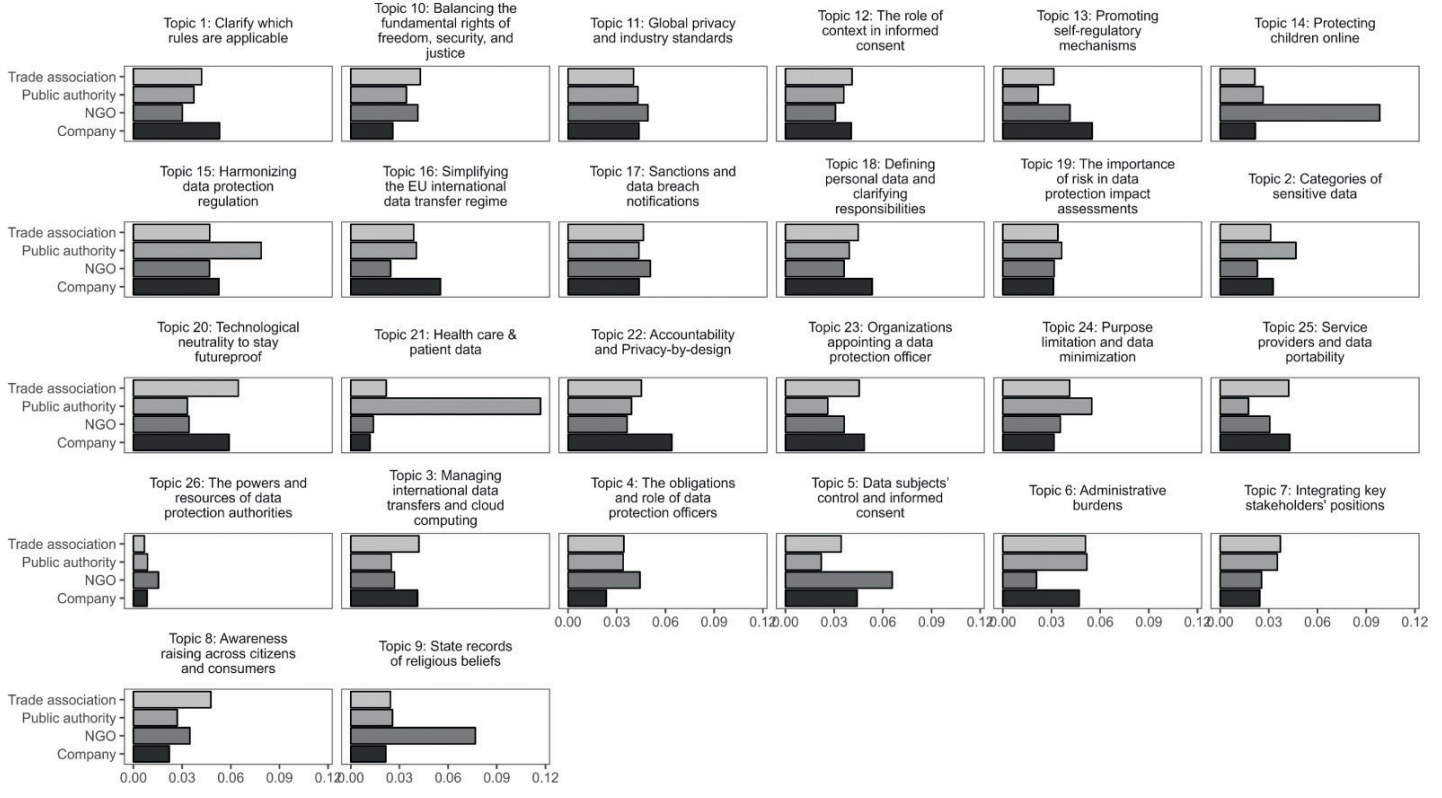


Figure 3.5. 26 topics from the "legal framework for the fundamental right to the protection of personal data" corpus with estimated author prevalence.



3.5 Discussion and Conclusion

The literature on CPA has done much to advance our understanding of how businesses lobby governments in pursuit of favorable policy outcomes. In this domain, scholars distinguish three specific approaches to corporate political strategy: financial incentives strategy; information strategies; and constituency-building strategies (Hillman & Hitt, 1999). Within the context of the EU, information strategies are central to achieving lobbying success (Coen et al., 2021). Yet, CPA scholars have little to say regarding the substance of information strategies (Lenway et al., 2022), except insofar as they aim to mold the public discourse (Nyberg & Murray, 2020). Furthermore, CPA scholars have generally eschewed the EU as their empirical context in favor of the U.S. (Chalmers & Macedo, 2021), and more recently emerging markets (Mbalyohere & Lawton, 2018). Given the fact that the EU often acts as a global norm-setter in areas such as digitalization, consumer health and the natural environment (Bradford, 2020), there exists an impelling need for CPA scholars to explore this relatively uncharted territory, not only to broaden the field's horizons but also to synergize with political scientists and public policy experts in understanding the dynamics of broader lobbying processes.

Responding to calls for such research (e.g., Lenway et al., 2022), we have explored the substance of information strategies within the context of the GDPR. Although we acknowledge the pioneering efforts of previous empirical studies on the GDPR (Andrew & Baker, 2021; Atikcan & Chalmers, 2019; Goyal et al., 2021; Kalyanpur & Newman, 2019), our contribution stands apart. Our study is the first to systematically and comprehensively explore the entire set of lobby documents submitted to the EC in response to two GDPR consultation rounds. Aided by advances in the field of unsupervised machine learning (Roberts, Stewart, & Tingley, 2014), we were able to uncover the latent themes in GDPR lobby documents.

Our study makes three important contributions to the CPA literature. First, we conceptualize two distinct information strategies. Initially, in the context where an official communication or position paper is non-existent, corporations alongside other stakeholder factions gravitate towards a ‘categorical information’ strategy. We delineate this approach as: “The act of imparting specific knowledge or perspectives to policymakers, seeking to persuade them to either wholly integrate or dismiss particular components

within prospective legislation." This strategy emerges as the predominant force in the early stages of EU lobbying, especially when it becomes obvious that legislation is on the horizon, albeit cloaked in ambiguity regarding its definitive contours. This pattern was manifestly evident in our dataset. For instance, we saw how during the first consultation rounds discussions focused on broad topics such as the international transfer of personal data (topic 17); principles and definitions, role of self-regulation/co-regulation (topic 16); and public interest and balancing conflicting fundamental rights (topic 7). We further illustrate this with proper examples from all author types in *Table 3.4*.

Subsequently, once an official communication is disseminated, companies and other stakeholders pivot towards a 'conditional information' lobbying strategy. We conceptualize this approach as: "The act of imparting specialized knowledge or insights to policymakers with an intention to persuade them to embed specific caveats or constraints within the impending legislation." We uncovered multiple examples in our topic model how during the second consultation round, lobbying efforts shifted towards more specific issue areas, such as the organizational appointment of a data protection officer (topic 23); the categories of sensitive data (topic 2); and accountability and privacy-by-design (topic 22) (see *Table 3.5* for examples). In summation, we submit that this bifurcation will equip CPA researchers with a pivotal instrument to delve deeper into the intricacies of information strategies, which, until now, remained a notably elusive domain in the CPA discourse (Lenway et al., 2022). We summarize the main components of both information strategies in *Table 3.4*.

Additionally, our study accentuates the need to broaden the understanding of institutions within nonmarket strategy scholarship, specifically by emphasizing the pivotal role of institutional procedures. Contemporary CPA research has been profoundly shaped by the foundational works of Baron (1995a, 1995b). Baron's theoretical constructs predominantly casts institutions as governmental agencies, exemplified by entities like Congress and the Federal Trade Commission. Yet, our findings illuminate a wider spectrum of institutional constraints that businesses must navigate. Specifically, throughout the policymaking process, corporations encounter enhanced prospects to mold public policy *prior* to regulatory entities crystallizing and broadcasting their legislative aspirations. These institutional procedures are critical for firms to recognize and incorporate in their

nonmarket strategies.

Finally, we also demonstrate that stakeholder groups take into account the lobbying activity of others and engage with this lobbying activity in subsequent rounds. This contradicts earlier research that found that interest groups do not in fact engage in counter lobbying (McKay & Yackee, 2007). Our finding seems to suggest that interest groups have come to the realization that they can achieve more from their lobbying activities by monitoring the behavior of their opponents.

In addition, on a methodological note, our study exemplifies the evolving frontiers of research methodologies in the CPA domain. Echoing the sentiments of Lenway et al. (2022), we underscore the immense potential of harnessing emerging technologies and innovative methodologies for more nuanced and sophisticated explorations into the intricate world of corporate political activities.

While our study offers valuable insights into the dynamics of business lobbying, it is not without its limitations, some of which pave the way for further exploration in the realm of CPA. To begin with, the clandestine nature of business lobbying is well-recognized: companies are naturally reticent about disclosing intricate details of their interactions with policymakers (Coen et al., 2021). Despite our best endeavors, the empirical data we acquired stems solely from submitted responses to the EC's consultation rounds, as per our FOI request. This inevitably raises the suspicion that a significant proportion of lobbying actions remain veiled, taking place away from the public eye and without official documentation.

Further, our empirical case is limited to a single piece of legislation within the context of a political entity that has been described as a 'sui generis' (Wallace et al., 2015). This means that the generalizability of our findings beyond the EU setting is unclear. However, we note that the 1946 Administrative Procedure Act requires all U.S. agencies to publicly publish all proposed rules and solicit comments from the public before moving towards enactment (McKay & Yackee, 2007), which means that further study to replicate or nuance our findings in the U.S. context would be possible. In addition, we recommend exploring other legislative areas in the EU context, such as the natural environment and consumer health. EU regulation in these areas has strong extraterritorial implications (Bradford, 2020), and thus constitutes an essential part of international firms' nonmarket environment,

also worth follow-up research.

In addition, we were not able to determine the efficacy of different information strategies, as this was beyond the scope of our study. However, future research can build on the concept of categorical and conditional information strategies and explore the degree to which each are effective in achieving their objectives, and whether this effectiveness differs between different actor types. For example, which actors' categorical information strategies were successful in prompting policymakers to include their recommendations as an anchor in their communication or draft proposal? Relatedly, did their conditional information strategies successfully yield desired caveats or exemptions on draft proposal? We suspect, however, that such research will have to combine various natural language processing (NLP) techniques, alongside domain expertise, to effectively address this research question.

In concluding, we wish to reflect on the some of the ethical intricacies of deploying NLP techniques on textual data, especially when dissecting lobby documents. First, the essence of FOI requests is to foster transparency, yet there looms the peril of uncovering sensitive information that might not have been meticulously redacted. This paves the way for inadvertent disclosures. The precision of NLP could potentially amplify such lapses, bringing forth genuine apprehensions about privacy infractions. Consequently, it becomes imperative for scholars to rigorously vet any information prior to its inclusion in publicized studies. Second, the nuance and contextual fabric of the original document can sometimes be overshadowed or misconstrued during NLP processing, yielding interpretations that may inadvertently misrepresent the original intents or positions of the stakeholders. This poses a tangible threat to the credibility and public image of the organizations in question. It becomes incumbent upon research teams, therefore, to immerse themselves in extensive, holistic dialogues to truly grasp the context in which these documents were written.

Table 3.4. Differentiating Categorical and Conditional Information Strategies

Information Strategy	Definition	Objective	Context
Categorical Information	The act of imparting specific knowledge or perspectives to policymakers, seeking to persuade them to either wholly integrate or dismiss particular components within prospective legislation.	Include or exclude wholesale element in forthcoming legislation	Regulators have not communicated their exact intentions; no existing stakeholder demands.
Conditional Information	The act of imparting specialized knowledge or insights to policymakers with an intention to persuade them to embed specific caveats or constraints within the impending legislation."	Introduce limitations or exceptions to expected elements in forthcoming legislation	Regulators have communicated their intentions; stakeholders have openly communicated their policy preferences

Table 3.5. 3 illustrative topics from the "comprehensive approach on personal data protection in the European Union" corpus, including example document by author type

Topic	Author			
	Company	***NGO***	***Public Authority***	***Trade association***
International transfer of personal data	The rules on model clauses for international data transfers are a good example of this phenomenon. Model clauses concern the existence – on paper – of rules that may not be observed in practice. All too often companies sign the clauses without putting in place appropriate mechanisms for ensuring their continued application in day to day life. Considerable time and energy is often expended on deploying model clauses that might be better utilised on ensuring an appropriate outcome – i.e., the protection of the relevant data in the country of importation.	However, Directive 95/46/EC did not expressly take account of BCRs. As a result the process for adoption of BCRs, which is based on Article 26 (2) of Directive 95/46/EC, requires the approval of all Member States concerned by a BCR. As a result, assessing and approving BCRs takes a long time. The WP29 has devoted considerable effort to promote and facilitate the use and the approval of BCRs within the current legal framework. In order to improve the process, so far, nineteen DPAs have agreed to a procedure on the approval of BCRs called 'Mutual Recognition	However, Directive 95/46/EC did not expressly take account of BCRs. As a result the process for adoption of BCRs, which is based on Article 26 (2) of Directive 95/46/EC, requires the approval of all Member States concerned by a BCR. As a result, assessing and approving BCRs takes a long time. The WP29 has devoted considerable effort to promote and facilitate the use and the approval of BCRs within the current legal framework. In order to improve the process, so far, nineteen DPAs have agreed to a procedure on the approval of BCRs called 'Mutual Recognition'.	Binding Corporate Rules (BCRs) represent a promising approach to international transfers. Such rules establish a unified global company standard, requiring accountability on behalf of the company for data protection regardless of where data are processed. As such, it is better adapted to the modern trends. We welcome the work undertaken by the Article 29 Working Party in developing a BCR toolkit, consisting of an application form, checklists and template for BCRs, and FAQs. We also welcome the development of a mutual recognition procedure which has been adopted by a majority of

Topic	Author			
	Company	***NGO***	***Public Authority***	***Trade association***
				Member States.
Principles and definitions, role of self/co-regulation	The Data Protection Directive (95/46/EC) (the "Directive") has established one of the most rigorous systems for data protection in the world. The EU should be commended for developing this robust framework, which is widely respected and which has served as a model for several other jurisdictions. The Directive also draws on international principles, such as the Fair Information Principles and the OECD principles, which contributes to a more harmonised global approach to data protection.	The accountability model envisages a regulatory environment that is far less based on prescriptive administrative procedures, but rather creates mechanisms to provide confidence that organisations processing personal information can be relied upon to ensure defined and accepted privacy outcomes if they are effectively held to account. The model is in the main based upon internationally agreed general privacy principles, the ones that already form the foundation of the Directive, which we discussed at the beginning of this paper. Industry specific Codes of Practice will play a greater role in specifying how the rules should be implemented in	Directive 95/46/EC should serve as a benchmark for the comprehensive framework which has as main goal effectiveness and effective protection of individuals. The existing principles of data protection need to be endorsed, and complemented with measures to execute these principles in a more effective manner (and to ensure a more effective protection of citizens' personal data).	BSA believes that the legal framework established by the Data Protection Directive, with its broad principle-based approach, use of common concepts, definitions and appropriate exceptions has enabled the application of Europe's data protection framework to remain flexible, technologically neutral and adaptable since 1995. To ensure the Directive remains effective, the BSA would welcome further discussions with the Commission on possible options that could be considered to address the identified challenges. Regardless of how the Commission chooses to proceed, we encourage the Commission to consider the following issues, and we look forward to discussing

Topic	Author			
	Company	***NGO***	***Public Authority***	***Trade association***
		particular settings.		these points in more detail going forward.
Public interest and balancing conflicting fundamental rights	In addition, there is a tendency to elevate data privacy to the rank of an absolute right, which can in turn effectively negate protection to other fundamental rights including that to an effective remedy. Although a fundamental right, data privacy and the guarantee of the application thereof "cannot be absolute and must yield on occasion to other legitimate imperatives such as the prevention of disorder or crime or the protection of the rights and freedoms of others". This requirement to balance fundamental rights is indeed unambiguously reflected in various important legal instruments as well as case law.	We are sure that the general principles of the protection of personal data have to be applied in this field also for ensuring the rights every EU citizens; therefore it is necessary to extend the scope of the new legislation to the fields of police and judicial cooperation in criminal matters also. Furthermore we think that it would be necessary to provide the suitable guaranties of using new technologies for data processing for these purposes (e.g. processing the DNA and fingerprint for law enforcement purposes or using video surveillance equipments for the purpose of crime prevention e.g.)	The justification of further data processing for the purpose of public causes is a matter of fundamental interest, which goes beyond the weighing of interests for the purpose of personal data protection. The fundamental right to personal data protection is not an isolated right, but should also be considered in its relationship to other fundamental rights guaranteed by the Charter of Fundamental Rights of the European Union. The right to liberty and security (Article 6 of the Charter) and the right to respect for private and family life (Article 7 of the Charter) also deserve protection. Government and law enforcement should take care of that.	In its judgment "Promusicae vs Telefonica", the European Court of Justice ("ECJ") stated, for the first time, that Member States in transposing the various directives on intellectual property, e-commerce and data protection, must strike a fair balance between the fundamental rights that they protect - including the right to property in civil proceedings - and must respect general principles of Community law, such as the principle of proportionality. The conclusion reached by the ECJ sent a strong signal to all stakeholders that neither data protection nor IP protection should be given precedence over the other.

Table 3.6. 3 illustrative topics from the "legal framework for the fundamental right to the protection of personal data" corpus, including example document by author type.

Topic	Author			
	Company	***NGO***	***Public Authority***	***Trade association***
Organizational appointment of a data protection officer	In addition, even where a company determines that appointment of a data protection officer is the best way to meet the requirements, under the accountability principle, that company should have the flexibility to determine whether to appoint a data protection officer (1) for each legal entity in a Member State where the company operates, (2) for all legal entities in that member state, or (3) for the whole of EU, depending on how that company organizes its business and privacy compliance activities internally.	For organizations that do opt to appoint Data Protection Officers, they will also need the flexibility to structure their data protection organizations in ways that will suit their specific industry, organizational or business model requirements. Some may want to designate individual data protection officers in each of their different lines of business or rely on a corporate-wide privacy team consisting of legal, HR and marketing experts from the individual business units to oversee compliance. Mandating the structure of data protection compliance without regard to size or industry sector will do little to encourage better privacy compliance.	The Norwegian DPA would like to share its experience regarding the value of Data Protection officers. In the Norwegian Privacy act the appointment of a data protection officer is not mandatory, but is initiated by the entity that wishes to establish such an internal function. This is appropriate in Norway because there are many small and micro enterprises that will be too small to have the economic strength to uphold such an independent position to fulfil the intentions behind a DPO.	As regards making the appointment of an independent Data Protection Officer mandatory, we have no objections to any company choosing voluntarily to appoint a Data Protection Officer (DPO). We however do not support any obligation for controllers to have a DPO, nor the imposition of mandatory policies and mechanisms for compliance with data protection rules. Such an obligation would be, in particular for small and medium-sized companies, an unjustifiable organisational and financial burden. However, we believe that controllers should have the possibility to appoint an internal or external DPO. In such a case controllers should then be freed from

Topic	Author			
	Company	***NGO***	***Public Authority***	***Trade association***
				the obligation of conducting filings with the DPAs.
Categories of Sensitive Data	We can see the case for genetic data being classified as sensitive data, but believe that any further proposals for additions to the category should be considered very carefully in the context of the motivation for the category. For instance, in the UK it is often noted with surprise that financial data is not considered "sensitive" for the purposes of the Data Protection Act. This is because the layman's view of sensitive data tends to be something like "data whose misuse (and particularly whose loss) would be particularly damaging for the data subject". Since the easiest way to quantify damage is in pecuniary terms, obviously financial data tends to appear high on people's lists of sensitive	With regard to sensitive data, the picture is also complicated. Member States may augment the definition of sensitive data beyond the data types included in the Directive. For example, sensitive data in Poland are defined as "addictions" and in Portugal as "private life." In contrast, in Spain, information relating to "lifestyle" (a notion similar to "private life") is explicitly regarded as non-sensitive data. There are also other inconsistencies. In Lithuania, ethnic origin is understood to mean "nationality"; in Spain "ideology" is understood to include political opinions; in Luxemburg there are various categories of health data while other States are less specific and only mention health.	Whilst all health information is rightly categorised as 'sensitive' data, recognition should be made of the fact that the notion of sensitivity can vary from one individual to another dependent on the circumstances. In some contexts for some people certain data might be deemed more sensitive than others. Therefore there can be a mismatch between what the law says and what people believe to be sensitive. As noted earlier in our response under 2.1.1 consideration should be given to the question of the context in which data is held and processed and of the potential risk of harm to the individual.	BSA believes that the legal framework established by the Data Protection Directive, with its broad principle-based approach, use of common concepts, definitions and appropriate exceptions has enabled the application of Europe's data protection framework to remain flexible, technologically neutral and adaptable since 1995. To ensure the Directive remains effective, the BSA would welcome further discussions with the Commission on possible options that could be considered to address the identified challenges. Regardless of how the Commission chooses to proceed, we encourage the Commission to consider the following issues, and we look forward to discussing these points in more detail

Topic	Author			
	Company	***NGO***	***Public Authority***	***Trade association***
	personal data.			going forward.
Accountability and privacy-by-design	Cisco believes that any legislative proposals relating to privacy-by-design should focus on encouraging flexible application of the principle to their design processes as opposed to introducing prescriptive requirements or specific technology mandates.	Some of these new principles and mechanisms for integrating privacy considerations into business models and product development cycles are referred to as "Privacy by Design." The concept has been prominently championed by Ann Cavoukian, Ontario's Information and Privacy Commissioner, and was endorsed by the Privacy Commissioners at their recent meeting in Israel. The concept encourages companies to take responsibility for data processing by inculcating a culture of privacy throughout their organizations. The Commission should explore ways to incentivize companies to implement "Privacy by Design" into the development of their offerings.	Public authorities have failed to provide examples of best practice in the EU. From health databases to smart metering and publicly-funded transport ticketing systems, there has been a fundamental failure to ensure proportionality, privacy by design and data minimisation.	We support the principle of Privacy by Design, which already guides our members' development processes. BSA would, however, appreciate further clarification from the Commission regarding how it proposes to define and integrate the principle of Privacy by Design into the data protection framework. We believe that Privacy by Design should be understood as a process for ensuring that data protection is carefully considered in the design and implementation of products and services. If this principle were instead used as a basis for imposing design mandates on particular technologies, it would hinder, rather than promote, user privacy and security.

APPENDIX TO CHAPTER 3
Probabilistic Topic Modeling

Probabilistic topic modeling rests on three foundational assumptions. First, it posits that each document comprises an amalgamation of various topics or subject matter. Second, it conceives of each topic as a constellation of words. This means that when a particular subject matter is discussed, a particular basket of words is used. Third, it assumes that the concurrent presence of words within a text is indicative of an underlying topic, or a 'latent structure'. Within this framework, a 'topic' is understood as an aggregation of words, each characterized by a quantifiable probability of association with that topic. At the 'top' of the topic are words that frequently occur within that topic, and thus achieve a relatively high probability. Conversely, a document is envisioned as a composition of multiple topics, wherein each topic is attributed a specific probability of occurrence within the document. While the documents themselves and the words they contain are directly observable, the topics, as well as the distributions of both topics per document and words per topic, remain unobserved. These constitute the 'hidden structure' of the dataset (Blei et al., 2003), which is revealed by the algorithm. Documents, here, can be anything from social media posts, lobby documents submitted by MNEs to the EU, or academic articles.

Topic modeling operates by leveraging the observed co-occurrence of words to infer this concealed structural arrangement in the form of topics. The emergent topics offer a reduced-dimensional representation of a corpus, facilitating applications such as classification and thematic description (Blei, 2012; Hannigan et al., 2019). The methodology of topic models is underpinned by probability distributions, whereby words are allocated to topics based on their likelihood of relevance to a topic, and the prevalence of a topic within a specific document (Blei et al., 2003). The term 'corpus' is employed to denote a compilation of text from a singular primary source, whereas 'corpora' refers to a collection of texts derived from multiple sources (Hannigan et al., 2019).

Latent Dirichlet Allocation (LDA) is a sophisticated technique for topic modeling that discerns the latent thematic structure within a corpus by deconstructing documents into two distinct matrices: the topic-per-document matrix (θ matrix) and the word-per-topic matrix (β matrix). This process essentially uncovers the inherent structure of documents by approximating the

resolution of a matrix factorization problem. Let M represent the entire set of documents, V the complete vocabulary, and K the aggregate of topics.

$$[M * K] * [K * V] = [M * V]$$

The matrix on the right side, known as the document-term matrix, is a sparse term count matrix characterized by one row per document and one column per term (Roberts et al., 2014). This matrix serves as the input for the LDA algorithm. The θ matrix quantifies the proportion of words in a document emanating from each of the K topics. Conversely, the β matrix encapsulates the natural logarithm of the likelihood of encountering each word given a specific topic. These matrices, as outputs of the algorithm, illuminate the thematic composition of the entire document collection, offering a nuanced understanding of the underlying thematic landscape.

The efficacy of topic modeling, particularly through Latent Dirichlet Allocation (LDA), lies in its ability to reconstruct the generative process of document creation. This reconstruction is achieved by feeding data into the LDA algorithm, based on the three foundational assumptions previously discussed. As a generative model, LDA provides a plausible explanation for the data generation process by constructing this explanation using a combination of various distributional building blocks. A crucial aspect to grasp is that certain fundamental components of this explanatory model are concealed; these hidden elements are inferred through probabilistic inference techniques (Blei et al., 2003).

The meaningfulness of the output generated by LDA can be attributed to its inherent design, which mandates a balance between two competing objectives: limiting the distribution of terms across a minimal number of topics within each document, and within each topic, assigning a high probability to a limited set of terms. This balancing act effectively identifies clusters of words that frequently co-occur, thus uncovering coherent topics within the data (Blei, 2012). This trade-off is central to the model's ability to discern and delineate the thematic structures embedded within the corpus. In other words, the algorithm is not overly generous in assigning a word to many different topics, and within each topic it identifies, LDA aims to focus on a relatively small group of words that are critical to that topic. The focus on this trade-off ensures that topics become quite interpretable, as the top words of

each topic should, in theory at least, are both unique and meaningful representations of higher-level subject matter. This balance is key to its ability to pick out and clearly define different themes from a large collection of documents.

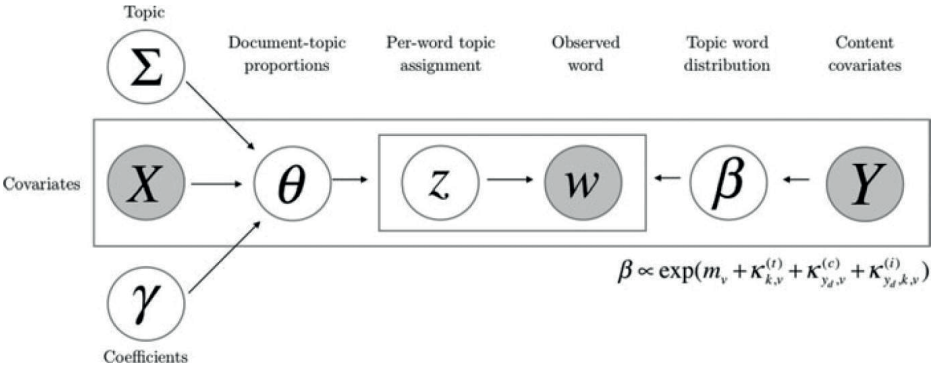
Hannigan et al. (2019, p. 590) highlight three principal advantages of generating topics through statistical probabilities in topic modeling. First, this approach liberates researchers from the need to impose preconceived structures on the documents they analyze. By not requiring a priori assumptions about document organization, topic modeling allows for an unbiased exploration of the data. Second, it enables the discovery and comprehension of significant themes that might elude human readers. This feature is particularly valuable in revealing underlying patterns and connections in large text corpora that might not be immediately apparent through traditional reading and analysis methods. Third, topic modeling accommodates polysemy, the phenomenon where a single term can have multiple meanings (*running* a marathon is not the same thing as *running* for public office), because it does not treat topics as mutually exclusive entities. In this framework, individual terms can be distributed across various topics with differing probabilities, allowing for the natural overlap and clustering of topics (DiMaggio et al. 2013, p. 578). This aspect of topic modeling is crucial for capturing the nuanced and multifaceted nature of language, as it acknowledges and represents the complexity inherent in the use of terms across different contexts.

The foundational work of Blei et al. (2003) on LDA has spurred further advancements in the LDA algorithm to enhance model accuracy (Blei & Lafferty, 2007; Roberts et al., 2016; Roberts, Stewart, & Tingley, 2014; Roberts, Stewart, Tingley, et al., 2014). A notable limitation of the original LDA is its incapacity to capture *topic correlations*. This addition makes topic models richer by taking advantage of the fact that discussing a certain topic increases the likelihood that a related topic is discussed within the same document. For instance, a document discussing datacenters is intuitively more likely to also mention cloud computing. Blei and Lafferty (2007) identified that under the Dirichlet distribution, the components of the proportion vector are almost independent, leading to the unrealistic assumption that topics are not interrelated. In other words, the inherent design of the Dirichlet distribution is inconsistent with the notion that particular topics are likely to

co-occur within documents. To overcome this, they introduced the correlated topic model (CTM), which employs a more flexible distribution for topic proportions, allowing for correlations among topics. CTM thereby provides a more nuanced representation of topic relationships, acknowledging that the presence of one topic can influence the likelihood of another's occurrence. The CTM is underpinned by the expectation–maximization (EM) algorithm (Blei and Lafferty 2007).

Structural topic modeling (STM), developed by Roberts et al. (2014), extends the capabilities of topic modeling by examining not only the relationships between topics and documents but also between arbitrary metadata, namely the documents' author(s) and the content and topics they produce. As such, STM's primary computational innovation is the incorporation of document-related metadata into the topic estimation process. This approach enables the examination of how various metadata elements correlate with both topic prevalence and content. Topic prevalence refers to the association between document covariates and the average proportion of a document discussing each topic, while topic content assesses the association between document covariates and specific word usage within a topic (Roberts et al. 2016). Put differently, the STM algorithms allows us to estimate whether certain authors are associated with particular topics, and whether certain authors are more likely to use particular words when discussing particular topics. *Figure 3.6* provides a visual representation of the STM framework.

Figure 3.6. A Visual Representation of Structural Topic Modeling



Source: Roberts et al. (2016, p. 990)

The distinctive feature of STM compared to traditional probabilistic topic models lies in its incorporation of covariates X and Y, as seen in *Figure 3.6*. These covariates significantly enhance the model's utility by establishing connections between topics and specific external variables, a crucial aspect for validating the effectiveness of the topic model (Grimmer & Stewart 2013). In this dissertation, STM's capability to ascertain whether a particular source of a document is linked with certain topics is particularly valuable, as will be demonstrated in Chapter 3. This estimation of correlations between document sources and topics contributes to a more precise model fit, as it facilitates the alignment of specific topics with their relevant document sources. Likewise, it allows us to scrutinize the model more profoundly by ascertaining whether a topic is discussed by a single author (such as a trade association) or multiple authors. This estimation process can be executed through various methods that integrate uncertainty (Roberts et al. 2014, 2016). In our application of STM of lobby documents in the GDPR, we incorporate diverse metadata into the model, a choice driven by the model's ability to provide a more comprehensive and accurate depiction of stakeholders' demands from the EU and the connections among these demands.

The initial phase in topic modeling, regardless of what particular algorithm is implemented involves the preparation of the corpora, a process pivotal to the integrity and utility of the subsequent analysis (Hannigan et al. 2019, p. 11). This stage includes selecting and refining the comprehensive set of documents, which significantly influences the outcomes of all ensuing analytical procedures. As with any algorithm, the quality of the output is determined by the quality of the input: garbage in, garbage out. For this reason, it is crucial for researchers to describe in great detail the choices they made at this stage, as it will have profound consequences for the final model.

Text preprocessing encompasses several key steps: eliminating extraneous white space, converting all text to lowercase, removing stop words (commonly used words with minimal semantic value), and implementing word stemming or lemmatization (Kobayashi et al. 2018). Additionally, words that appear infrequently across the document set may also be excluded, as recommended by Roberts et al. (2014). This not only streamlines the dataset but also helps in discarding documents of lower quality. Word stemming involves reducing words to their base or root form, for instance, transforming 'management' to 'manag'. Lemmatization is a more sophisticated process

where words are converted to their base or dictionary form, so that ‘regulations’, ‘regulatory’, and ‘regulated’ all become ‘regulate’. Some researchers argue that this is beneficial for model accuracy as it consolidates nearly identical words, thereby enhancing the model's capacity to detect closely associated word clusters. However, this believe is contested by computer scientists (Schofield et al., 2017), who found that this common practice has no measurable effect on the quality of the topic model, and in some circumstances can either hurt the model’s interpretability. Furthermore, the elimination of prevalent custom terms within the document set is advised, as these can skew the estimation of topics (Silge & Robinson, 2017).

The subsequent phase in topic modeling involves deciding on the optimal number of topics k for the algorithm to estimate. Following the preprocessing steps, this is the second critical decision for researchers to make. This decision is guided by two key computational metrics: semantic coherence and exclusivity. Semantic coherence, a concept developed by Mimno et al. (2011), is predicated on the idea that in models exhibiting high semantic coherence, the most probable words under a topic are likely to be found together within the same document. Models are considered more semantically coherent when their coherence scores are closer to zero. Exclusivity, on the other hand, can be quantified using the FREX metric, as proposed by Airoldi and Bischof (2016). FREX stands for the weighted harmonic mean of a word’s rank according to its exclusivity and frequency. Thus, a best practice approach involves selecting models that score highly on both these measures. However, this process requires a balance; models with fewer topics tend to have higher semantic coherence but may lack in exclusivity for more numerous topics.

Despite the value of semantic coherence and FREX metrics in topic modeling, it is vital to recognize that topic modeling serves primarily as a tool to enhance the clarity of textual data analysis. A topic model, for social scientists, is a lens we use to examine large amounts of textual data in a comprehensive and meaningful manner. Thus, the determination of the optimal number of topics in a topic model is not a question with a definitive answer, as noted by Hannigan et al. (2019) and Roberts et al. (2014). In more technical terms, the selection of the number of topics to estimate is less about achieving precise population parameters and more about identifying the most illuminating perspective to understand the data, as emphasized by DiMaggio et al. (2013, p. 582). The goal is to uncover and render the hidden thematic

structures within the corpora in a way that is comprehensible to human interpreters. As such, it is advised that researchers estimate various models with different preprocessing decisions and number of topics, and engage in systematic comparisons to determine the optimal model. If the objective of the research question extends to assessing the impact of specific metadata on topical prevalence and content, an additional step in structural topic modeling involves constructing functions for the covariates X and Y. In Chapter 3, which aims to gain insights into the substance of lobby documents of firms, trade associations, public authorities and NGOs, we introduce the document's author as a covariate for topic prevalence.