The Anatomy and Evolution of ESG Reports

Ethan Rouen*
Harvard Business School

Kunal Sachdeva
Rice University

Aaron Yoon
Northwestern University

October 2021

Abstract

We study the anatomy and evolution of ESG reports, examining how firms and standards influence the decision to make voluntary disclosures on ESG issues. Using a hand-collected sample of all ESG reports for S&P 500 firms from 2010-2020, we analyze their content, document how these disclosures have changed over time, and identify factors that influence the disclosure choice (e.g., leading firms and voluntary standards). Using topic modeling algorithms guided by the Sustainable Accounting Standards Board’s ESG standards, we provide evidence for the influence of other firms and the influence of these standards in the dynamic shaping of voluntary ESG reports.

JEL Classification: M14, M40
Keywords: ESG Reports, SASB, Voluntary Disclosure, Textual Analysis, Topic Modelling

*Corresponding author contact information: erouen@hbs.edu. We are indebted to Janine Guillot, Nicolai Lundy, and Arturo Rodriguez at the Value Reporting Foundation for providing access to SASB-related data and feedback. We thank Jim Naughton, Bob Kaplan, Joseph Pacelli, Krishna Palepu, George Serafeim, Rodrigo Verdi, and workshop participants at the Harvard Business School for their thoughtful comments. We acknowledge the excellent research assistance of James Enright, Kunal Rai, Wesley Ramos Dos Santos, Mei Tercek, and Irpan Yierpan.
1. Introduction

This study provides the first large-scale examination of firm Environmental, Social, and Governance (ESG) reports. ESG reports are among the fastest growing voluntary disclosure phenomena in recent history. At the start of the 21st century, almost no large, publicly traded U.S. firms released these reports, but 20 years later, a large majority of the firms in the S&P 500 have published stand-alone ESG disclosures. ESG reports have become more detailed, providing information about firms’ impacts on a range of issues, from their carbon emissions to the diversity of their workforces to their charitable activities. Despite their growing frequency and complexity, firms and investors struggle with how to define material ESG activities, and ESG reports remain voluntary, meaning that not all companies report them and those that do can choose what they disclose. In addition, these reports describe largely unobservable activities and are unaudited, yet their content is of growing interest among stakeholders (Christensen, Hail, and Leuz 2018, henceforth CHL).¹

The lack of uniformity among ESG disclosures has led to at least two problems. First, it makes it difficult for investors, researchers, and stakeholders to parse the data in these disclosures in order to conduct time-series and cross-sectional analysis. As a result, even for-profit data providers have published conflicting ESG performance ratings for the same company, and it is difficult to understand how these providers arrived at different assessments.²,³ Second, there is significant heterogeneity across time and firm in what ESG issues are considered material, so

---

¹ As evidence of this growing interest, assets under management of signatories of the United Nations Principles for Responsible Investment grew from a few hundred billion dollars in 2006 to $120 trillion by 2020 (Rouen and Serafeim 2021).
² For example, Tesla is ranked as one of the highest and lowest ESG performers in different large data providers such as MSCI and Sustainalytics.
³ Data providers such as Bloomberg and Sustainalytics do not provide access to the underlying data used in their ratings, leading to an additional concern: that the black-box nature of these data makes it impossible for researchers and investors to understand the origins of the conflicting ratings.
creating standards is a challenge. Perhaps the most successful attempt to develop uniform standards is that conducted by the Sustainable Accounting Standards Board (SASB).⁴ In 2015, 85% of firms disclosed at least some information on a topic SASB deemed financially material in regulatory filings, and within this sample, firms disclosed 23 SASB-related excerpts, on average (CHL). Still, most of these disclosures were boilerplate, meaning that, at least within regulatory filings, companies provide little specific information of relevance to stakeholders (CHL).

Studying the content in ESG reports is, on its own, of interest to accounting scholars given the growing interest in this topic and the push both internationally and in the United States for requiring additional ESG-related disclosures in firms’ regulatory filings. Still, the rapid uptake in ESG disclosures by publicly traded firms also provides a unique setting to understand whether and how firms coalesce around a set of relevant topics in voluntary disclosures that are of interest to various stakeholders.

Therefore, this study also aims to contribute to the broad literature on voluntary disclosure by examining how firms influence each other’s reporting choices in the absence of regulation, and how voluntary standards are related to the dynamic shaping of disclosures. Specifically, we aim to (i) examine the type of content and topics included in voluntary ESG reports among S&P 500 firms; ii) provide a detailed examination of the evolution of these reports during the last decade; and (iii) study the extent to which industry leaders shaped the creation of SASB’s standards and how those standards shaped the evolution of ESG reports of other firms.

To accomplish this goal, we conduct a large-scale hand collection of ESG reports of all firms in the S&P 500 from 2010 to 2020.⁵ Because there exists no clearinghouse for ESG reports

---

⁴ We provide more detail about SASB below.
⁵ ESG reports may go under the name of corporate sustainability report or sustainability report. In this paper, we will refer to them as ESG reports.
comparable to the Security and Exchange Commission’s EDGAR website for financial disclosures, we hand-collect these reports from company websites, and supplement this strategy with searches of archival websites. With these data, we examine the evolution of these reports over time and how voluntary standards impacted the content in these reports. We ask the following three related research questions:

1. What information do firms choose to disclose and how has that content changed over time?
2. Do early disclosers shape standards and peer firms’ disclosure choices?
3. Do voluntary standards shape what firms disclose in ESG reports?

To explore question (1), we document the evolution of the uptake of ESG reports since they began being published in order to determine how this voluntary disclosure spreads across the market. Our goal is to provide descriptive evidence related to the decision to disclose and the content of those disclosures.

We begin by determining what drives the decision to voluntarily disclose ESG-related information in a separate report and identify potential market drivers that made these disclosures widespread. In this set of analyses, we implement a Cox proportional hazard model to measure the likelihood that firm releases an ESG report. Our event of interest, releasing an ESG report, is voluntary and thus is right censored. Given the nature of the question and event of interest, we use a semi-parametric approach. Because industry structures are different, we stratify our baseline hazard by industry codes. Further, we control for a host of firm characteristics that are plausibly predictive of firms’ propensity to disclose ESG activities, such as firm size, ownership structure, governance, and management characteristics in order to develop a determinants model (CHL).

---

6 We download additional reports from the GRI Disclosure Database and the Corporate Register.
To analyze the content of these reports and how they have evolved, we use a guided (i.e. semi-supervised) topic modeling algorithm, relying on SASB’s materiality maps as a framework when defining our topics of interest.\(^7\) These maps, which were introduced in a staggered manner between 2013 and 2016, provide sector-specific standards that were developed in partnership with industry working groups (IWG) consisting of investors, auditors, regulators, and companies (see Figure 1 for the sector-level materiality map).\(^8\) We discuss the maps and their evolution in detail in Section 2.2.

For our topic modeling approach, we provide seed words from the SASB standards for each topic as a starting point for a graph-based algorithm. From this starting point, our algorithm determines other associated words from other ESG reports to form topics. This approach allows us to identify and directly measure the amount of content devoted to financially material topics versus content unrelated to these material issues, resulting in a panel at the topic-firm-year level measuring the percent of each document that addresses each topic. For example, the algorithm determined that business ethics, a material topic in the Financials industry, comprised 17% of the content in Goldman Sachs’ 2020 ESG report.

Relying on our panel at the topic-firm-year level, we explore question (2) by studying how much firms that released ESG reports prior to SASB guidance addressed material topics. In this analysis, we focus on the set of early disclosing firms. Within these firms, we identify two subsets.

---

\(^7\) Researchers have commonly relied on unsupervised models (e.g. Latent Dirchlet Allocation or LDA) to model the topics of a corpus of documents. This is because there doesn’t exist a strong prior about what topics exist. While this approach is broadly used in the literature, a problem is that they result in unlabeled, and often, a diffuse set of topics. In further robustness tests, we implement a state-of-the-art BERT model and manually match the topics to the materiality map.

\(^8\) We note that SASB is not the only organization that gives guidance on material ESG issues. For example, Global Reporting Initiative (GRI) provides guidance on ESG’s materiality from stakeholders’ perspective. SASB, on the other hand, has a narrower focus on shareholders and provides guidance on the financial materiality of ESG issues. Given SASB’s focus, and that large companies and investors have begun relying on SASB to benchmark firms’ disclosures, our focus on the SASB materiality map also allows us to focus on ESG topics that are material to investors.
The first subset of firms is those that worked directly with SASB as part of its IWG to develop its standards. The second subset of firms is formed by the remaining early disclosing firms. Using the two groups of firms, we examine the amount of each ESG report that is related to each topic of materiality prior to the release of the preliminary SASB disclosure guidance.

We explore question (3) by studying the extent to which SASB standards shape firms’ voluntary ESG disclosures. This question is motivated by concerns that ESG disclosures are unaudited, and as such, firms can discuss in these reports either issues that are directly related to their operations or ancillary activities that may serve as little more than a marketing tool (i.e., “greenwashing.” See Delmas and Burbano 2011; Kim and Yoon 2020). For example, while the average firm that discloses at least one SASB-relevant topic in regulatory filings includes 23 ESG-related excerpts, only three of those excerpts are non-boilerplate or contain quantifiable metrics (CHL).

We use a difference-in-differences empirical specification to estimate the change in financially material disclosures around the release of SASB guidance. The SASB standards were developed with industry-specific guidelines (i.e., each industry has its own set of material topics). With these standards being released for each industry in a staggered pattern over time, we measure the increase in material disclosures, relative to non-material topics. We include firm, time, and topic fixed effects to address concerns of time trends, firm invariant characteristics, and the average level of disclosures for each topic.

In our prior analysis of question (2), we determined if and how IWG firms differentially disclosed material topics in the pre-period. However, what is yet to be determined is if, and how, such guidance differentially affected IWG firms and their peer firms. We study the heterogenous treatment effects of disclosures on IWG-firms by splitting our sample into IWG firms and other
early disclosers, and re-estimating our difference-in-differences specification. Uncovering the source of change (i.e., which subsample is driving the change in disclosure), in combination with our prior analysis of IWG firms in the pre-period, provides insight into whether and how early disclosers shape standards and peer firms’ disclosure choices.

These analyses allow us to provide evidence around two important issues. First, we document how firm disclosures can potentially shape regulatory standards. Examining both subsets of firms allows us to differentiate between the influence of early disclosure on its own and the influence of early disclosure among those that had a relationship with the guidance provider. Second, we provide empirical evidence on whether and how firms influence their peers’ disclosure choices in the absence of mandatory disclosure standards. While the evidence provided in these analyses is not causal, the setting provides a unique opportunity to document how reporting choices diffuse within an industry and identify some of the factors that can shape voluntary reporting and the standard setting process. Given that some firms released ESG reports prior to the release of the SASB standards, our attempt sheds light on the dynamic shaping of disclosure that arose around the release of these standards.

We are careful in how we interpret these estimates because it is well known that SASB developed its standards over time, and with the industry working group. Further, the areas of material topics were not chosen randomly, but instead, with industry-specific challenges in mind. As such, it would be incorrect to interpret our results as causal. That is, the random assigning of disclosure guidance will not necessarily change how firms disclose information. Instead, we interpret these results as being the treatment effect on the treated — the counterfactual would be if SASB developed its reporting guidelines but did not release these standards to the public.
Lastly, we provide early descriptive evidence to prove the viability of this proposal by conducting a pilot study focused on two sectors: Extractives & Minerals Processing (EMP) (i.e., non-renewables) and Financials. We chose these two sectors because of the differing characteristics in their ESG reporting. While many EMP firms have been reporting ESG-related issues for more than a decade, these disclosures were largely absent from financial firms prior to 2010. In addition, there is little overlap in the material ESG issues of these sectors, as defined by SASB, making them an appropriate selection in order to test our algorithms and provide descriptive statistics.

This paper makes several contributions to the growing body of literature examining ESG disclosures. Importantly, CHL provides a comprehensive review of the academic literature and discusses CSR disclosures’ capital market effects, real effects in firm behavior, and implementation issues related to the adoption of CSR standards. In addition, it uses U.S. regulatory filings collected by SASB in 2015 and documents a substantial heterogeneity in CSR disclosures, both across as well as within sectors. We extend CHL by examining nearly 10 years of voluntary ESG disclosures and documenting the evolution of voluntary disclosure of ESG issues, in addition to shedding light on the role that standard setters and peers (i.e., industry leaders) play in the evolution of these disclosures. In addition, we analyze the content of these voluntary reports, which adds to CHL’s examination of regulatory filings. This strategy allows us to analyze the role of voluntary standards on voluntary disclosures and better understand the details of these disclosures. We believe the findings will inform regulators across the globe as they consider making some ESG-related disclosures mandatory.

While ESG reports as a phenomenon are of a great interest and importance on their own, they also provide a window into how disclosure spreads through a market, given their novelty and
ubiquity. A large body of research has examined the spillover effects of firms’ disclosures on their peers (e.g., Foster, 1981; Baginski and Hinson 2016; Capkun, Lou, and Wang 2019; Breuer, Hombach, and Muller 2021). Much of this research has focused on the proprietary costs of disclosure and how firm disclosure impacts the demand for peer firms’ disclosures (Baginski and Hinson 2016; Capkun et al. 2019). Benefitting from the rapid adoption of ESG reports, we take a different approach and document how the content in firms’ voluntary disclosures is influenced by the disclosures of peers, whether ex ante voluntary disclosures influence how organizations such as SASB provide guidance to the market, and how these voluntary standards influence voluntary disclosures. More broadly, we examine how voluntary ESG disclosure evolves over time.

Finally, we contribute to the recent literature on the reliability of ESG ratings by documenting and quantifying the actual disclosure content that shapes the ratings. There has been much disagreement among ESG rating providers on how to define firm ESG performance, which caused a significant amount of confusion in the capital market (Chatterji, Durand, Levine, and Touboul 2016; Berg, Koelbel, and Rigobon 2020; Serafeim and Yoon 2021). Currently, commercial data providers (e.g., MSCI, Sustainalytics, and Bloomberg) often use voluntary disclosures and other information available to them (most of which are derived from firm-initiated disclosures), while exercising discretion and judgement to rate firms’ ESG efforts. The result is often an aggregate score that is opaque to data users, making it difficult, if not impossible, to back out how the ratings are determined. Our paper takes an important first step towards understanding what may shape the discrepancy because we quantify ESG disclosures upon which many of the ESG ratings are at least partly based. We plan to make the data and all relevant codes available publicly. This process will provide other researchers with a dataset of machine-readable measures of ESG disclosures that are intuitive and interpretable, which in our view, may help our
understanding of the discrepancy between ESG disclosure and vendor-initiated ESG performance scores.

The rest of this proposal proceeds as follows. Section 2 provides institutional details about the setting. Section 3 discusses the data, the pilot study, and variable measurement. Section 4 describes the empirical methodology. Section 5 concludes.

2. Institutional Setting

2.1 ESG Reports

As a subject, ESG in business has been a fast-growing phenomenon in recent years and much attention has been paid to it by investors and companies. For example, signatories of the United Nations Principles for Responsible Investment (PRI) had only a few hundred billion dollars in assets under management (AUM) in 2006, but the AUM reached $120 trillion by 2020. This growing salience of ESG is not unique just to the asset management industry. There were fewer than 20 publicly listed companies in the world that issued reports that included ESG data in the early 1990s. By 2014, the number of firms reporting on ESG issues had increased to nearly 6,000 firms globally (Serafeim 2014). In the United States, 83% of companies registered with the Securities and Exchange Commission (SEC) disclose some sustainability information in their regulatory filings (SASB 2017).

Initially, firm-initiated disclosure of sustainability information started from press releases and disclosures on company websites when firms were thrown into the spotlight after high-profile scandals and events (CHL). Such disclosure practices became recognized as industry best practice and served as guidelines. In response to growing demand for ESG information from investors, stakeholders, and regulators, firm-level ESG disclosures began being centralized in one venue:
ESG reports. Against this backdrop, a stream of academic literature examined the firm-level determinants of sustainability reporting (see Hahn and Kuhnen 2013 for a detailed review).

Many of these studies are either small sample studies and/or use settings outside of the United States, but they raise concerns about whether ESG reports contain information that accurately reflects firm ESG performance. Dilling (2010) finds among European companies that those in the energy sector and those with higher profit margins were more likely to produce GRI-compliant ESG reports. Papers such as Simnett, Vanstraelen, and Chua (2009) and Manetti and Becatti (2009) document that firms seek voluntary assurance on their non-financial reports to enhance credibility, but there is great variation in whether firms seek assurance, the amount of information reported, and the validity of that information (Perego and Kolk 2012). Relatedly, Roca and Seary (2012) analyzes Canadian companies in 2008 and finds that there are more than 500 different indicators in sustainability reports. Boiral (2013) examines 23 sustainability reports that received an A or A+ from the GRI and finds that most of the firms’ negative ESG events were not reported, suggesting that the voluntary nature of these reports leads to selective disclosure. Similar in spirit, Hubbard (2011) analyzes sustainability reports of 30 of the world’s largest companies and finds that they lack discussions related to targets and performance.

This concern is also reflected in recent literature that highlights how data vendors consume ESG information and produce ratings. As of 2016, there were more than 100 ESG data providers (Amel-Zadeh and Serafeim 2018), relying on voluntary disclosure and other publicly available information to create ESG ratings (Chatterji et al. 2016; Serafeim and Yoon 2021). These ratings are heavily subjective, and there are differences in how data vendors define, weigh, and measure ESG, which causes significant confusion in the investment community as to what ESG entails (Berg et al. 2020).
2.2 Sustainability Accounting Standards Board

There are at least two organizations that offer voluntary reporting guidelines for ESG activities. Their aim is to improve or harmonize reporting practices in response to the demand for information and the current inconsistent state of corporate ESG disclosures. The two most prominent organizations that provide guidance on materiality of ESG information are SASB and GRI.

SASB was founded in 2011 as a nonprofit organization, and develops and disseminates sustainability accounting standards to help publicly listed corporations voluntarily disclose material factors in compliance with SEC requirements. SASB standards are designed for the disclosure of financially material sustainability issues in mandatory SEC filings, such as the Form 10-K and 20-F. GRI was founded in 1997 and requires that a GRI-compliant report cover issues that reflect the organization’s significant economic, environmental, and social impacts; or issues that substantively influence the assessments and decisions of stakeholders. The key difference between SASB and GRI is that SASB has an investor focus and GRI has a multi-stakeholder focus.

We are guided by SASB’s standards in this analysis because it provides several unique advantages for our setting (i.e, the U.S. market) and for our time period. First, because of SASB’s focus on identifying and connecting the link between ESG issues and shareholder value, it has been better received by U.S. listed firms than has GRI. Second, SASB’s staggered adoption of standards gives us a tighter setting to identify the impact of early disclosers’ disclosure practices on the shaping of standards as well as the subsequent impact on other companies. Third, SASB provides sector specific guidance on the financial materiality of ESG issues. Of course, materiality in theory would vary at the firm-time level. However, this financial materiality lens at the sector
level allows us to observe in detail how disclosure practices of a firm may have shaped the disclosure practices of its industry peers.\textsuperscript{9}

The SASB’s board comprises a mix of regulators, academics, lawyers, and investors. Its standards are developed via a multi-stakeholder process consisting of research supported by Bloomberg technology, data, and analytical tools; balanced, multi-stakeholder industry working groups; a public comment period; and reviews by an independent Standards Council of experts in standards development, securities law, environmental law, metrics, and accounting.

SASB initially had 79 provisional industry standards across 10 sectors. These were published sequentially by sector between July 2013 and March 2016. Figure 2 presents the timeline of the SASB standard setting process. In the initial research phase, SASB collects evidence for each industry of the financial impact of sustainability issues to identify the industry-specific materiality of sustainability activities and related metrics. Unlike most industry classification systems that use sources of revenue to group companies into different sectors and industries, SASB uses a Sustainable Industry Classification System (SICS) to group similar companies based on their sustainability-related risks and opportunities.

Then, industry working groups of stakeholders are organized to provide feedback on the identified issues and metrics. These groups have balanced representation from corporations, market participants, and public interest intermediaries. Their feedback on the materiality of topics and usefulness of metrics is incorporated into the exposure draft standard.

In the final phase, the exposure draft standard is released for a 90-day comment period for any member of the public to provide feedback. Feedback is then analyzed and incorporated into a provisional draft standard. The standards were considered final when the complete set for all

\textsuperscript{9} As disclosure of ESG issues are now of a global interest, SASB and GRI announced in November 2020 their intention to merge into a unified organization, the Value Reporting Foundation by mid-2021.
industries was completed and reviewed by the American Standards Institute, an independent Standards Council comprised of experts in standards development, securities law, environmental law, metrics, and accounting.

SASB published sector-level provisional standards and sought public review from July 2013 through March 2016 (see Appendix A for sector-level dates of adoption). After incorporating feedback through January 2018, SASB released in November 2018 a finalized set of codified standards for 77 industries across 11 sectors. The materiality guidelines span five dimensions: environment; social capital; human capital; business model and innovation; and leadership and governance (see Figure 1 and https://www.sasb.org/standards/archive/ for more details).

3. Data, Variable Measurement, and Pilot Study

In this section, we describe our methodology for collecting and parsing ESG reports, as well as the variables we create. As a proof of concept and to provide descriptive evidence of the implementation of this methodology, we also conduct a pilot study. We include in this section a description of the pilot study sample and descriptive statistics from the pilot study.

3.1 ESG Reports Data

This paper’s analysis centers on the ESG reports of all firms in the S&P 500 from 2010 to 2020. Because there is no centralized database of reports, we conduct a large-scale hand collection of ESG reports for these firms.\textsuperscript{10} We chose the S&P 500 for this initial approach as this sample includes only large firms that have the necessary resources to disclose. In order to create a consistent time series, we include all reports for all firms that were included at least once in the

\textsuperscript{10} For our pilot study, described below, we downloaded and analyzed all ESG reports for all S&P 500 firms in the EMP and Financials sectors.
S&P 500 during our sample period, resulting in more than 700 unique firms. Upon the completion of this study, we will make this data and code fully available as a public good for other researchers and practitioners.

As a starting point to conduct an exhaustive search for ESG reports on all companies, we downloaded all available reports from firms’ websites. When archival reports were not available, we supplemented our search by searching the GRI database. Finally, we conduct an exhaustive Google search of archival websites.

Because ESG reports are unstructured, not standardized, and in PDF format, accurately extracting text poses a significant challenge. To ensure accurate extraction, we used several machine learning techniques to extract the textual contents of the documents accurately. Specifically, we use the following process. As a first step, we iterated through each page of each document and saved the files separately, which allowed us to create an optical scan of each page, resulting in high-quality images. This strategy is superior to simply relying on commonly used libraries to directly identify the text in PDFs since PDF encoding introduces noise that results in excess and incorrect words. Next, we used an open-source optical character recognition (OCR) engine to extract the text from each page. Because the OCR process introduces errors in the documents, we used natural language processing (NLP) techniques to correct any misspelling in the text and remove any artifacts introduced in this procedure. Finally, we hand verified the extracted text to ensure that we accurately constructed a corpus representative of the ESG reports.

3.2 Pilot Study Sample

To provide evidence of the plausibility of this strategy and offer early descriptive figures, we conducted a pilot study, creating our variables (described below) for the two sectors. The goal
of this pilot study is not to provide empirical evidence to support our research questions but to demonstrate the computationally intense process of developing our variables and methodology.

We chose the EMP and Financials sectors because they are among the most different in our sample in terms of ESG reporting along two important dimensions. First, while EMP firms have been disclosing ESG-related activities for decades, few firms in the Financials sector made these disclosures prior to 2010. Figures 3 and 4 shows the uptake in standalone ESG reports for firms in our pilot sample. As reported in Panel B of Figure 3 (Panel A reports absolute numbers of ESG reports), fewer than 20% of firms released ESG reports in 2010, but almost 70% did so by 2020. Figure 4 examines uptake by sector, as well as SASB adoption by sector. While the financial sector has seen faster growth (from a lower initial number of disclosures), both sectors have experienced dramatic growth in the number of firms that provide ESG reports. In addition, almost no firms had adopted the SASB guidance by 2018, when the standards were codified, but since then, as reported in Panel A (Panel B), almost half (30%) of firms in EMP (Financials) have adopted the SASB guidance.

3.3 Using Machine Learning to Quantify Disclosures

We use several machine learning techniques to quantify the textual characteristics of ESG reports. Below we outline the two key measures, and describe this process in detail in Appendix B.

3.3.1 Measuring Topics

Our main empirical analysis relies on measuring how much of each ESG report is devoted to addressing each ESG topic identified by SASB. To accomplish this, we must develop a model
of the language used in the ESG reports and then iterate through each document to quantify the disclosures of each topic.

Given the questions asked in the paper and our strong priors about the context of the documents, we use a semi-supervised approach to model the topics contained in ESG reports. Our baseline approach uses a state-of-the-art graph-based model (see Appendix B.1 for technical details). The graph model is built upon a co-occurrence network at the token level where each node represents a token and an edge represents the occurrence of token 1 and token 2 within a certain range of words in the sentence. From this graphical representation, we provide our model with “seed” words that are associated with the topics that correspond to SASB’s materiality topics. Using the seed words as starting points, we use various techniques to traverse the graph to identify related words and form topics.

Using our model, we can determine groupings of words within each topic. Figure 5 plots the sample topics and words associated with the topics based on ESG reports from our pilot sample where a word’s relative size signifies its importance. Panel A shows the topic for air quality. We see related words, such as “ozone,” “carbon,” and “barrels” are associated with this topic, based on similar seed words. Similarly, Panel B shows the topic for data security. Here we see related words like “malicious,” “threshold,” and “cyber.” For these topics, it is important to note that we did not pre-define all the words, but instead, the words were identified by the algorithm.

The final dataset is a panel at the topic-firm-year level that captures the percent of each topic present in each document. In essence, we measure the relative prominence of each topic in

---

11 We test numerous models. For example, as an alternative approach, we use a bidirectional encoder representations, or BERT, model developed by Google in 2018. This model has the advantage of generating embeddings that allow us to have multiple vector representations for the same word. This results in context-dependent embedding. The main drawback from this approach is that it is unsupervised and results in unlabeled topics.
each document. Figure 6 presents a heatmap of the relative content of materiality for each topic. Panel A plots the average topics discussed for the EMP sector, while Panel B plots the same for the Financials sector. We can see that certain topics have become more prominent, while others have disappeared, over time. Topics with green bars are those that are material to firms in the sector, as defined by SASB, while red bars are those that are not material. Darker shades represent that these topics are discussed more frequently among firms in the sample. White bars represent topics that have relatively little content discussed in the documents, on average.

3.3.2 Computing the Similarity Between Documents

We compute a cosine similarity between each document and construct a matrix of this measure. This matrix, Similarity, allows us to study both the within-firm similarity over time (e.g., Goldman Sachs in 2014 and Goldman Sachs in 2015) and across-firm similarity (e.g., Exxon Mobil in 2015 and Chevron in 2015). Looking at within-firm Similarity for the pilot sample of EMP and Financials firms, mean (median) Similarity is 0.92 (0.95), meaning that firms’ current-year ESG reports contain much of the same information as the prior year’s report. Given the small sample of the pilot, it is difficult to interpret this statistic. On one hand, such similarity could signal boilerplate disclosures. On the other hand, it could suggest that firms have coalesced around meaningful disclosures that are presented in a similar way every year, similar to the 10-K. We leave analysis of this measure to the final paper. Still, to further explore this measure descriptively, we report in Figure 7 the year-over-year average Similarity across firms (within the same sector). The cross-sectional Similarity has significant variation over time. In 2010, the Similarity across firms was more than 64%, while in 2018, it hit a low of just above 58%.
4. Research Design

4.1 Identifying the Determinants of Publishing ESG Reports

We initially set out to identify which firms choose to publish ESG reports. However, a challenge that we face in studying this question is that of censoring, or the choice to disclose. The event in this paper, publishing an ESG report, is not always observed by the end of the study period for each firm. However, there is still valuable information in knowing that a company chose not to publish a report by the end of our study period. As such, our empirical strategy incorporates this information into its estimates.

The appropriate estimation method for the question in this paper is a hazard model. For our baseline specification we use a Cox proportional hazard model, expressed as:

\[ h_g(t,X) = h_{0g}(t) \exp \{ \beta'X \}. \]  

The first factor, \( h_{0g}(t) \), is the baseline hazard function, and is left unspecified, while the second factor, \( \exp \{ \beta'X \} \) is the shift factor, with the regressors entering linearly. The covariate is equal to zero. The shift factor equals one, and does not contribute to the hazard rate. It is called a proportional model because estimated covariates are assumed to affect the baseline hazard rate, \( h_{0g}(t) \), across the entire domain of time. This model can be further estimated by using stratifications for each industry, \( g = 1, \ldots, k^* \). Stratification may be appropriate, as the baseline hazard function, \( h_{0g}(t) \), may be different for each stratum specified in an estimation.

We consider a set of explanatory variables not limited to the following: Firm Size is the natural logarithm of the book value of assets. Market-to-book is defined as the market value of equity plus book value of debt over book value of assets. Returns is the stock return in the past 12 months. Profitability is the ratio of earnings before interest, taxes, depreciation, and amortization scaled by sales. The above variables are to account for the possibility that larger, more profitable,
and highly valued firms may have the capability to invest more and/or disclose more on ESG issues. *Institutional Ownership* is the percentage of outstanding shares held by institutional investors and is to account for the pressure to incorporate ESG from these investors. *Refinitiv Score* is the ESG rating of the individual firm provided by the Thomson Reuters database and is converted to a standardized score for ease of interpretation. This variable is to account for the possibility that firms with more ESG investments would be more inclined to disclose on their ESG endeavors.

4.2 *Do Early Disclosers Shape Standards and Peer Firms’ Disclosure Choices?*

4.2.1 *Did Industry Leaders Already Disclose More Material Topics?*

We next examine whether industry leaders disproportionately disclose in areas that SASB subsequently deemed to be material. A unique aspect of our research setting is that SASB solicited guidance from certain firms when crafting its standards. These firms, members of SASB’s industry working group (IWG), arguably had greater insights into the standard setting process and were able to anticipate standards before they were released (CHL). This aspect of the standard setting process allows us to ask whether early disclosers (i.e., members of the working groups) influenced the disclosure choices of other firms in the sector.

*Ex ante*, it is unclear if firms that were part of the IWG were disclosing in areas that were subsequently deemed to be material. On one hand, industry leaders may have provided guidance that was in line with their current practices, suggesting that there is a first mover advantage to help shape the SASB guidance. On the other hand, if the guidance was different from the IWGs’ ESG practices, it would suggest a learning-by-doing approach, where leading firms dynamically update their beliefs on what is material, and it is reflected in the guidance. Therefore, understanding what
industry leaders were disclosing prior to SASB’s guidance is important in mapping out if and how industry leaders affected future disclosures by non-industry leaders.

We study this question by comparing between firms that were releasing ESG reports prior to the release of SASB guidance, which we call early disclosers. Then, we form two subgroups. The first set of disclosing firms are those that were part of SASB’s IWG. The second group of firms were also early disclosures but were not affiliated with the IWG. Guided by our hazard model, we use a propensity score matching approach to identify counterfactual firms similar in terms of observable characteristics, including sector.

We create our firm-level variable by measuring the level of disclosure on each topic and adjusting for the industry average on the topic during year t. This is to understand where firm i stands relative to other firms in the same sector. We use the following equation:

$$HighDisclosure_{i,j,t} = 1 \left( \%Topic_{i,j,t} - \frac{1}{N} \sum_{t=1}^{N} \%Topics_{j,t} > 0 \right),$$  \hspace{1cm} (2)

where $HighDisclosure$ is an indicator equal to 1 if a firm has above mean disclosure on a topic, compared to other firms in the sector, and $i$, $j$, and $t$ correspond to firm, topic and year, respectively. With this measure of relative disclosure, we estimate the relation between membership in the IWG and the propensity to disclose material topics. Specifically, we estimate the following equation:

$$HighDisclosure_{i,j,t} = \beta_1 (IWG_{i,t} \times MaterialTopic_j) + \delta_i + \delta_j + \delta_t + \epsilon_{i,j,t}. \hspace{1cm} (3)$$

The coefficient of interest, $\beta_1$, measures the relative propensity of IWG firms to release disclosures in material topics. Estimating a positive coefficient would provide suggestive evidence that IWG firms were already producing disclosures that were meaningfully different from their peer firms. In contrast, if we are unable to reject the null of a zero estimate, it would point to early disclosers producing ESG reports not too different from other firms.
4.3 Do Voluntary Standards Shape ESG Reports?

4.3.1 Did Disclosures Converge Towards Topics of Materiality?

To better understand what shapes the materiality map, we engaged in multiple phone calls with SASB executives and researchers. Guided by our discussions, we design an empirical strategy to shed light on how the implementation of SASB guidance is associated with the discussion of materiality in ESG reports for all firms in our sample.

In this section, we examine the impact of voluntary standards on the shaping of disclosures in ESG reports. The sample is at the topic-firm-year level from 2010 through 2020. The topic-firm-year observations measure the percent each topic is discussed in a given firm-year report. Specifically, we estimate a difference-in-differences regression model:

\[
Topic_{i,j,t} = \beta_1 I(MaterialTopic_{i,j})I(Post_{i,j,t}) + \beta_2 I(Post_{i,j,t}) \\
+ \beta_3 \xi_{i,t-1} + \delta_i + \delta_j + \delta_t + \epsilon_{i,j,t} \tag{4}
\]

The dependent variable \(Topic_{i,j,t}\) is the percentage of a firm’s ESG report devoted to a certain topic, where the topic corresponds to the set of topics put forward by SASB, and differs for each industry. \(I(MaterialTopic)\) is a dummy variable and equals one if the topic is material for a given firm. The \(I(Post)\) variable is a dummy variable and equals one for the firm-year following the first implementation of the SASB guidance for the industry. We include firm fixed effects, \(\delta_i\), to control for time-invariant characteristics of the firm. We include topic fixed effects, \(\delta_j\), to account for the average support level of each topic. Further, we include time fixed effects, \(\delta_t\), to control for time trends within the data. The interaction coefficient, \(\beta_1\), is the primary variable of
interest and corresponds to the differential change in material topics from firms after SASB guidance, relative non-material topics.

Our null hypothesis is that firms will increase their disclosures differentially towards material topics, versus other topics. This would result in a positive coefficient on $\beta_1$ and suggest that ESG reports are being formed directly by SASB guidance.

4.3.2 Did Disclosures Change Prior to the Guidance?

After establishing the relation between disclosure choices and SASB standards, we turn to examining whether the quality of disclosures changes prior to the release of SASB guidance. Using a difference-in-differences approach, our underlying assumption is that topics would evolve in parallel prior to our event. Although we are unable to fully test the parallel trends assumption, we take steps towards validating our empirical approach.

To test the possibility that differential pre-trends between material and non-material topics do not drive these results, we estimate the following equation:

$$Topic_{i,j,t} = \sum_{k=-\tau}^{\tau} \lambda_k d[t+k]_{i,t} + \sum_{k=-\tau}^{\tau} \beta_k \{d[t+k]_{i,t} \times MaterialTopic_{i,j} \} + X_{i,t-1} \zeta + \delta_j + \epsilon_{i,j,t} \quad (5)$$

We use the percent each topic is discussed in each document as our dependent variable. $d[t+k]_{i,t}$ is an indicator variable and take the value of one if the firm is $k$ years away from receiving guidance by SASB. $MaterialTopic_{i,j}$ equals one if the topic is material, and zero otherwise. We also control for variables that we find important in determining a firm’s propensity to disclose.
Estimates from this equation are used to compare the dynamic coefficients around guidance. For the parallel trends assumption to hold, we would expect the $\beta_k$ coefficients in the pre-periods to be insignificant, suggesting no differential pre-trends. On the other hand, significant $\beta_k$ coefficients in the pre-period would suggest that firms are coalescing around material standards prior to the release of the standards. Such a result would be interesting on its own and would warrant further empirical investigation.

4.3.3 Do Disclosures Converge Within a Topic?

We next examine whether firms converge in the way that they discuss topics. While the prior estimation approach is informative on the intensive margin of disclosures around voluntary guidelines, how firms disclose within a given topic remains unanswered. Prior studies, such as CHL, have highlighted the lack of understanding of this question and the importance of answering this question. In this subsection, we lay out our approach to extend CHL and characterize the convergence of material disclosures within topics.

Our baseline approach to study the convergence of topics is to use the absolute distribution of words used within an identified topic and study the pattern over time. Using the original list of words from our topic model, we recalculate the occurrence of each word, instead of the weighted importance of each word, and tabulate the frequency. Studying each topic for each report, we compute the variance of word choices over time, and call this measure $Variance of Topic_{i,j,t}$. Using this new measure of convergence within a topic, we re-estimate equation (4) by replacing our dependent variable as follows:

$$Variance of Topic_{i,j,t} = \beta_1 I(MaterialTopic_{i,j})I(Post_{i,j,t}) + \beta_2 I(Post_{i,j,t}) + \delta_i + \delta_j + \delta_t + \epsilon_{i,j,t}. \ (6)$$
Similar to our prior analysis, we are able to study the heterogeneity between material and non-material topics around the release of SASB disclosure guidance.

4.3.4 Did Disclosures Change Differently for Those Part of the IWG?

What firms, if any, led to the differential change in material disclosures? While establishing the association between SASB guidance and material disclosures is important, it is equally important to understand which firms are driving this differential change following the release of guidelines.

We study the heterogenous treatment of firms by splitting our sample based on a firm's involvement with the IWG prior to the standards release and re-estimating equation (4). Splitting our sample into IWG participants and those that were not included, we consider if and by how much firms are changing their material disclosures post guidance. Estimating two sets of regressions, we can compare the coefficient of interest to determine if and where the changes in material disclosures are coming from (i.e., by the firms outside of the IWG or those within).

This test allows us to separate whether the differential changes in material disclosures are coming (i) from early disclosers that were part of the IWG, (ii) early disclosers that were not part of the IWG, or (iii) from both groups. Finding that the changes are coming primarily from (i) would suggest that the standards were primarily catering to the IWG and its intended path of disclosures. Finding support for (ii) would suggest that firms unaffiliated with the guidance were differentially affected by the guidance. Finding any result in this section would shed further light about the role that IWG membership had and its effect on disclosures.
4.4 Discussion of Potential Empirical Results

The findings in this paper are likely to be of interest to regulators, standard setters, firms, and academics. While largely descriptive in nature, the results will provide evidence about whether and how disclosures diffuse through markets in the absence of mandatory regulation. Findings that support the efficacy of voluntary standards could provide evidence of an effective way to guide firms toward a uniform set of disclosure, while findings that support the influence of early adopters may incentivize firms to disclose additional ESG-related information or provide more detailed and relevant reports. Still, it is unclear whether either of these mechanisms are an effective tool for creating common disclosures.

This concern is reflected in the current debate about the usefulness of commercial ESG ratings databases, and their conflicting ratings. Given our research design, it is difficult to make definitive causal statements, but finding that neither market leaders nor voluntary standards influence reporting choices would be an important outcome. It would help to explain the many conflicting results in ratings and the academic literature (e.g., Berg et al. 2020; Christensen et al. 2021; Serafeim and Yoon 2021), given that similar firms are disclosing different information and different raters are making different recommendations on a topic even when they assess the same set of disclosure, making comparability a challenge.

This outcome would also be of use to regulators such as the SEC, which is currently examining whether to mandate disclosures related to environmental impact and human capital. For example, environmental related information generally has higher correlation as these items are often quantifiable metrics (e.g., CO2, GHG emissions) while issues related to the social aspects of ESG (e.g., employee satisfaction and firm culture) have significant amounts of disagreement (Welch and Yoon 2020).
5. Conclusion

In this paper, we study the anatomy and evolution of ESG reports, which have become important and ubiquitous voluntary disclosures in the last decade. Despite their growing frequency and complexity however, ESG reports remain voluntary, meaning that not all companies report them, and that those that do can choose what they disclose.

Against this backdrop, we propose to examine how firms and standards influence the decision to make voluntary disclosures on ESG issues. We will analyze a hand-collected sample of all ESG reports for all S&P 500 firms from 2010-2020, and analyze their content, document how these disclosures have changed over time, and identify factors that influence the disclosure choice (e.g., leading firms and voluntary standards).

In a pilot study, we use topic modeling guided by the Sustainable Accounting Standards Board ESG standards and provide descriptive evidence on the EMP and Financials sectors. We also plan to provide evidence for the influence of other firms and the influence of these standards on shaping voluntary ESG reports. Finally, we plan to make our data and code publicly available for academics and practitioners.
References


FIGURE 1: Sector-Level Materiality Map

Dark (light) grey color means that for more (less) than 50 percent of the industries within the sector, the issue is material. White means that the issue is not material for any industry within the sector. To see materiality maps at the industry level, visit: [https://materiality.sasb.org/](https://materiality.sasb.org/). The labels under each issue are generic. This means that the substance of the issue can differ dramatically from one industry to another. For example, supply chain management appears as material for both pharmaceutical and iron ore steel producer firms. However, in the case of pharmaceutical companies, “Manufacturing and Supply Chain Quality Management” refers to “Description of FDA enforcement actions taken in response to violations of current good manufacturing practices (cGMP), including: product deemed adulterated, form 483s, suggested recall (Class I, II, III), Warning Letters, Border Alerts, license suspension or revocation, product seizure, Consent Decrees, criminal prosecution. Description of corrective actions implemented in response to actions,” and to “Percentage of facilities and Tier I suppliers participating in the Rx-360 International Pharmaceutical Supply Chain Consortium audit program or equivalent third-party audit programs for integrity of supply chain and ingredients (e.g., APIs, chemical, raw material, excipients, etc.).”
FIGURE 2: Generic timeline of SASB standard adoptions

This figure displays the process through which SASB develops its standards. While the figure reports each step that every sector undergoes in order to reach codified standards, the actual dates and amount of time between steps differ at the sector level. Provisional standards were published from 2013-2016, while codified standards for all sectors were finalized in 2018.
FIGURE 3: Companies Increasingly Presenting ESG Reports

This figure presents the growth in companies presenting ESG reports within our pilot from 2010 to 2020. The horizontal axis is for the year the ESG report was released. Panel A presents the number of ESG reports by year. The vertical axis represents the tabulated count of ESG reports in a given year. Panel B presents the percent of companies within the S&P 500 presenting ESG reports. The vertical axis is computed by measuring the number of companies reporting divided by the total number of firms in the S&P 500, within our pilot.
FIGURE 4: ESG Reports and SASB Standard

This figure plots the increase in ESG reports for the pilot sectors of this study. Panel A is for extractives & minerals processing, while panel B is for the financials sector. For both plots the vertical axis represents the percent of firms reporting in the S&P 500, within a given sector. The solid black line is the sector side disclosure of ESG reports. The dotted grey line are the percent of firms that attest to following SASB standards. The dashed red line draws attention to the release year of the final release date of the SASB standards.
This figure provides a visual representation of important words detected by our topic mode. The relative size of the word represents the importance of the word within a given topic. Panel A represents the Air Quality topic from the SASB materiality map. Panel B represents the Data Security topic from the SASB materiality map.

FIGURE 5: Topics of Materiality

(a) Air Quality

(b) Data Security
This figure provides a heatmap of the relative content of materiality for each topic. Panel A plots the average topics discussed for the extractives & minerals processing sector, while Panel B plots the financial sector. From these plots, we can see that specific topics have become more prominent, while others have disappeared. Topics with green bars are those that are material to firms in the sector, as defined by SASB. Topics with red bars are those that are not material. Darker shades represent that these topics are discussed more frequently among firms in the sample. White bars represent topics that have relatively little content discussed in the documents, on average.
FIGURE 7: Cosine Similarity of Text Across Firms

This figure plots the textual similarity of ESG reports across firms. The errors bars plot the 95% confidence interval around the sample mean.
## Appendix A: Timeline of Sector-Level Standards Release

<table>
<thead>
<tr>
<th>Standards</th>
<th>Key Events</th>
<th>Key Dates</th>
<th>Extractives &amp; Minerals Processing</th>
<th>Financials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standards Outcome Report</td>
<td>12/12/2013</td>
<td>6/27/2013</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Exposure Draft of the Provisional Standards</td>
<td>1/14/2014</td>
<td>11/15/2013</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Public Comment Letters (Provisional Phase)</td>
<td>4/22/2014</td>
<td>11/19/2013</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Responses to Public Comments</td>
<td>7/1/2014</td>
<td>12/30/2013</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Provisional Brief</td>
<td>6/24/2014</td>
<td>2/1/2014</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Final Provisional Standard</strong></td>
<td><strong>6/1/2014</strong></td>
<td><strong>2/1/2014</strong></td>
<td></td>
</tr>
<tr>
<td>Development of Codified Standards</td>
<td>Technical Agenda</td>
<td>7/28/2017</td>
<td>7/28/2014</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Basis for Conclusions for the Proposed Changes to the Provisional Standards</td>
<td>10/1/2017</td>
<td>10/2/2017</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Exposure Draft of Proposed Changes to the Provisional Standard</td>
<td>10/1/2017</td>
<td>10/2/2017</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Public Comment Letters (Codification Phase)</td>
<td>1/31/2018</td>
<td>1/31/2018</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Summary of Public Comments</td>
<td>1/31/2018</td>
<td>1/31/2018</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Final Codified Standard</strong></td>
<td><strong>10/1/2018</strong></td>
<td><strong>10/1/2018</strong></td>
<td></td>
</tr>
</tbody>
</table>
Appendix B: Topic Modelling

Appendix B.1: Graph Approach

Our baseline approach uses a graph model that is built on a co-occurrence network in token level where each node represents a token and an edge represents the occurrence of the token 1 and token 2 within a certain range of words in the sentence. The co-occurrence graphs have been explored in different ways within the NLP field to determine relationships between entities such as defining a word embedding representation for non-ascii symbol [1] and snowball method to expand the search on initial terms to a new seed set [2, 3).

Initially, a window variable has to be defined as the look-around window for each token. To build the network, for each token i, there will be a list of co-occurred words within [i - window: i + window] where each combination with the word i is an edge of the graph.

The model can define topics in two ways. The first one is by performing a breadth-first search from a seed word or a set of words. This search will expand the area around the seed words to the new nodes where all the starting words are connected. This search returns a new set of words that can be further expanded until we reach a maximum number of iterations or there is no new word to be added to the topic without breaking the rules of co-occurrence. The second way to search in the graph is by finding densely connected regions. The algorithm implemented on this second method is based on the Louvain method for community detection [4].

The output for the first search method is a list of words that are highly related to the initial seed set, expressing mainly the context where we can find those tokens. Nodes that cannot be reached by any starting set can be interpreted as outliers and have no or very low relation to the seed words. The second method returns a list of nodes that can be interpreted as highly connected
communities or topics. To visualize the topics, word clouds with inverse term frequency as weight are plotted and metrics regarding the distribution of topics on the original files are stored.

To determine seed words with more than one token, it is necessary to split them into individual tokens and search for new tokens that co-occur with both of them. This way, we can perform topic modeling at n-gram level.

This method is extremely flexible. However, it is limited to the token or small sentence level with no more than 3 or 4 tokens. Also, this model is highly dependent on the initial seed sets and the context where they will be. A poorly defined seed set can lead to a mix of topics coming from the same search or even the impossibility to expand after the first iteration.

Appendix B.2: BERT Model

Our unsupervised topic model relies on a Bidirectional Encoder Representations from Transformers (BERT) approach dedicated to finding topics at the sentence level. For every text in the dataset, an overlapping split is applied to create smaller chunks with some overlapping to preserve specific sentences. Each chunk is transformed to a token-like vector with BERT embedding. The list of vectors is used to determine topics by performing two searches, a k-means approach and a hierarchical density-based spatial clustering method. In both scenarios, it is possible to determine the percentage of each document that belongs to each topic and the corresponding sentence. In the final paper, we will supplement our analysis using this approach to provide evidence that our results are not sensitive to the choice of algorithm we use.

The BERT model is an offshoot of the model proposed in [2]. It uses the context of a sentence-level encoding to transform small chunks of the document into a sentence that can be represented as a position in an n-dimensional space. The method combines this representation with a clustering approach to determine regions where sentences are categorized based on their meanings.

Each document is divided into smaller pieces of text with size n that contains some overlapping from the previous sentence. Each text is encoded using a BERT transformer [1], producing a list of encoded sentences for each document. More details about the contextual embeddings used are presented in [4].

The input size for the clustering algorithm is defined by the size of the embedding model. In order to reduce this size, the UMAP [3] algorithm is applied to reduce the number of dimensions. This approach leads to a faster convergence and minimizes the appearance of outliers.
There are two clustering models that we make use of. The first is a hierarchical density-based clustering model (HDBSCAN) where each sentence is grouped as part of a community or an outlier [5]. This method is set with a minimum of 7 sentences per cluster and euclidean distance to minimize. The second model is the k-means algorithm with a finite number of centroids that are randomly started and uses numerous iterations to form clusters.

The first method does not limit the amount of communities, which can lead to a high number of topics. To tackle this problem, the number of topics can be reduced by merging two or more into one based on cosine similarity until it reaches a feasible size.

The output on both methods is a list of sentences with the respective cluster. To visualize the topics, word clouds with inverse term frequency as weight are plotted and metrics regarding the distribution of topics on the original files are stored.

In this approach it is possible to determine topics in sentence level, as well as occurrence of n-grams of tokens within the topic.

This model does not require any initialization regarding the topics that we can find. At the same time this is useful because the approach does not require knowledge about the topics. However, it is more challenging to define or give any guidance to the model about specific topics. Even if it is possible to look for the topics after the clustering process, it is not possible to prevent them from being split along the clustered or even being put together with others. Also, this model is sensitive to the size of the sentences, and this value can be challenging to define properly, mainly when one document can have many topics that are discussed only briefly.


