International Journal of Disclosure and Governance (2020) 17:73–85
https://doi.org/10.1057/s41310-020-00087-w

ORIGINAL ARTICLE

# Exploring the U.S. Securities and Exchange Commission's Edgar database by sampling joint venture contracts

**Andres Velez-Calle**[1] · **Cristina Robledo-Ardila**[1]

Received: 5 August 2019 / Published online: 28 July 2020
© Springer Nature Limited 2020

## Abstract

Through its electronic filing system EDGAR, the U.S. Securities and Exchange Commission (SEC) provides information that listed companies are required to report. However, while the information is thus made public, it is not readily available. This article explains the purpose of the SEC and its EDGAR filing system, how it works and how anyone can access this information. It further explains how companies file their material contracts as attachments to their filings and how to find them. The paper also explains the development of a program for mining EDGAR, resulting in the extraction of over six hundred JV contracts. Finally, the paper explores basic text mining techniques to dissect the contracts and point out the most widely used clauses, such as joint venture type and scope, the laws governing the contract and the description and responsibilities of the parties and opens a broad research agenda.

**Keywords** Alliance governance · EDGAR · Securities and Exchange Commission · Data mining · Web scraping · Contract disclosure · Joint venture contracts

## Introduction

US publicly traded companies are required to disclose their information by filing their financial reports and corresponding attachments with the *U.S. Securities and Exchange Commission (SEC)* using its electronic system known as EDGAR.[1] This information is available to everyone, including investors who use it to make decisions. Even so, accessing it may require not only prior knowledge of the EDGAR database, but the use of new technologies and data mining techniques. In recent years, researchers in the fields of accounting and finance have increasingly consulted this database, but less so in the law and management fields. Companies themselves do not seem to use it to its full potential given a generalized lack of knowledge about the possibility for accessing information that is disclosed and updated for free. In consequence, they turn to commercial databases that offer paid, yet more user-friendly, access to limited information. This article provides a guide that will allow anyone, including researchers, managers or financial analysts, to access and use, in a systematic manner, information contained in public records for thousands of global companies that are required to report to the SEC.

The creation of algorithms to retrieve data from the SEC has become increasingly popular (Garcia and Norli 2012a; Loughran and McDonald 2017), considering that a relatively simple program can be used to download any firm's quarterly or annual reports later than 1996. Along with such reports, companies must file other documents relating to their financial activities and results, including contracts. This article uses joint venture agreements to illustrate the search strategies and technical aids available to improve the process of retrieving, organizing and analyzing information in a systematic manner to achieve meaningful and useful results. For this purpose, the paper shows how to find such contracts in the EDGAR database and suggests ways to create algorithms to download them to a spreadsheet. Broad findings relating to a first dissection of joint venture contracts are also presented.

To date, the study of the microfoundations of joint ventures lacks detail. The use of actual contract agreements allows exploring such detail and provides improved understanding of alliance design and governance, among other aspects. The contract is where the parties specify the control

Andres Velez-Calle
avelezca@eafit.edu.co

1 Universidad EAFIT, Medellin, Colombia

[1] https://www.sec.gov/edgar/searchedgar/companysearch.html.

and coordination mechanisms that govern the alliance and is the means for specifying what is allowed, penalties for breaching the agreement and expected outcomes and performance of the alliance (Argyres et al. 2007; Gong et al. 2007). Even though contracts provide details of the functioning of the alliance, it is never possible to anticipate all future contingencies; thus they are by nature incomplete (Hart and Moore 2007). To prepare for possible future situations, the parties use contracts to establish substitutive—or complementary—governance mechanisms that provide the alliance with flexibility and timely decisions throughout its execution. Such mechanisms include, but are not limited to, the share of equity and representation on the board of directors. The study of joint venture contracts and their anatomy contributes to a better understanding of ways to deploy governance structures that maximize joint venture value creation, especially considering that the structural governance of an alliance is key to its ability to add value through its alignment with the objectives (Sampson 2004). Also, it allows improving preparation and negotiation processes by identifying key areas that may help increase JV success (Luo 2005).

The article is organized as follows: "Literature review" section provides a brief literature review relating to the U.S. Securities and Exchange Commission and EDGAR. "Material contracts of firms" section reviews the main aspects of joint venture contracts and explains their availability through EDGAR. Also, the search procedure for contracts, along with key challenges, is illustrated for joint venture contracts. "Dissecting and analyzing the JV contracts" section summarizes key findings related to this type of contract, followed by a discussion section. Finally, conclusions are drawn and practical implications and directions for future research are suggested.

## Literature review

### The Securities and Exchange Commission (SEC)

The SEC was created by means of the Securities Exchange Act of 1934 with the purpose of restoring the trust in the financial system lost due to the stock market crash of 1929 and the Great Depression (Bushee and Leuz 2005). Through an extension of this law in 1964, the SEC required all public companies to file and publicly disclose their quarterly (10-Q) and annual reports (10-K) and accompanying documents (Gerdes 2003). At the time, the task of manually processing such a vast amount of information quickly became overwhelming. Access, even though electronic storage was available, was only possible in five public locations across the USA or through a few private companies that charged a fee. The creation of the Electronic Data Gathering, Analysis, and Retrieval System (EDGAR) marked a milestone in the process of effectively filing and making public firm information online (Gerdes 2003; Griffin 2003). Nevertheless, some obstacles remain.

### EDGAR

Voluntary electronic filings were first implemented in 1984 and became mandatory in 1993 (SEC 2000), resulting in improved filing and accessibility to reports and accompanying documents. The online electronic filing and retrieval system known as EDGAR was broadly applied in January 1996 (Griffin 2003). Since then, public firms have increasingly filed reports and documents online, which become publicly available online and free of charge almost instantly. Within approximately 40 s, records become available for everyone; however, *Tier 1* subscribers may receive the information a few seconds in advance, providing them with a trading advantage (Rogers et al. 2017). Nowadays, EDGAR is a massive repository of corporate and financial information on global US publicly traded companies containing millions of records. It has become a first-source firm repository and one of the richest sources for free and updated information of its kind today (Loughran and McDonald 2017).

### Types of filings and documents

The types of documents most commonly filed through EDGAR include Form 4 (changes in ownership), followed by the 8-K (earning releases), SC 13G/A (ownership of stock over five percent) and 10-Q (quarterly report) forms (Garcia and Norli 2012a). Refer to Garcia and Norli (2012a) for details on frequencies of forms filed in the EDGAR database. Investors and researchers mostly use the 10-K and 10-Q filings, which offer detailed and accurate information relating to the financial information of firms and are therefore considered the most important and widely used for investment decisions (Griffin 2003). Form 4 is also considered relevant as changes in the ownership of firms may signal shifts in their growth potential.

Regarding research fields, EDGAR filings have been widely used in finance and accounting, and investor reactions to 10-K and 10-Q filings can be divided into pre- and post-EDGAR eras. Studies in the pre-EDGAR era provided weak evidence on the relationship between timing and investor reaction (Easton and Zmijewski 1993). However, post-EDGAR era researchers can access electronic data shortly after filings are made, thus making the use of SEC filings more extensive and the accounting and finance literature using SEC filings more prolific. For example, scholars have found that on 10-K and 10-Q filing days, there is an increase in firms' stock trades (Griffin 2003). Regarding 10-K form filings, it has been found that those firms that are listed on larger stock exchanges, which in turn have in place stronger auditing and higher analyst coverage, are more likely to file late (Dalton et al. 2013). Also, You and Zhang (2009) found that investors tend to underreact to the timing by paying increased attention to the filing and its complexity, especially considering that it contains nonfinancial information that could predict future performance and investor reaction. Other studies in the post-EDGAR era have focused on different parts of the filing and their effects on company valuation. This is the case of De Franco, Wong and Zhou (2011), who analyzed the market reaction to the information provided in the notes to financial statements reported in 10-K filings, as these can be used by equity analysts to update financial statements, which in turn affect stock value.

The use of textual analysis in accounting and finance is an emerging area and is normally applied in the form of targeted phrases, similarity measures among documents, sentiment analysis or topic modeling (Loughran and McDonald 2016). The availability of computational methods to extract large amounts of information from databases has allowed an increasing use of this data analysis technique to examine information from varying sources, including that available through Exchange Commission (SEC) filings. More recently, advances in both data mining and text analysis have allowed the extraction of text from SEC filings, which in turn permits deeper analysis of content, creating a robust stream of literature relating to 10-K reports. Several approaches have been proposed for this purpose, including the template-based approach by Cong, Kogan and Vasarhelyi (2007) for extracting financial data from unstructured SEC filings. Others, including Loughran and McDonald (2011), used a predetermined H4N tag[2] dictionary of positive and negative words as a template to analyze the total sentiment of annual reports by classifying the positivity or negativity of their wording, also known as sentiment analysis. The authors added a new negative word list to the dictionary and then related the sentiment of the annual report to firm performance. Others, such as Garcia and Norli (2012b), used automated programs that extract geographical dispersion of operations data from 10-K forms to relate the geographical location of companies to their stock returns. They found that US firms with operations concentrated in one or two states obtained higher returns than those that were more geographically dispersed.

EDGAR has been improving since its inception, not only in terms of the accuracy and timeliness of information available, but also in terms of its search and retrieval mechanisms. Nevertheless, it is almost impossible to extract data from the enormous number of contracts and their attached documents using the provided EDGAR search engine. Every single company required to file reports through EDGAR is identified by a central index key (CIK) number. In principle, documents should be uploaded using a header indicating the type of filing (e.g., 10-K, 10-Q), and tags to highlight parts of the filings to give them similar structure (Gerdes 2003). But in most cases, companies fail to follow these guidelines, which may be due to ignorance, negligence or even as a deliberate strategy. In consequence, searching and retrieving information from EDGAR can be frustrating. For instance, the use of search queries is complicated and does not permit within text searches, which means that the search is limited to the headings provided—or not provided—by the filing company. In the case of documents attached to filings such as material contracts, the matter becomes even more difficult as these are filed using an irregular pattern, making them unsearchable. The lack of structure, organization and systematization within the filing of reports and associated documents is partly due to the complexity of the EDGAR's interface and is illustrated by the growing number of paid sources that offer user-friendlier databases, such as *Rankanfiled,*[3] *SEC info*[4] *and SECFilings.*[5]

As illustrated in this section, in the post-EDGAR era, access to SEC filings has become increasingly possible. Though in the early 2000s, annual and quarterly reports had to be downloaded manually one by one (Griffin 2003), nowadays new technologies allow access to thousands of reports using programs known as scrapers or spiders (Garcia and Norli 2012a). Research is now expanding to include the attachments to these filings, such as contracts, which are harder to locate because of misfiling. For instance, merger and acquisition contracts are meant to be filed using form S-4 but are commonly attached to other filing types (Sanga 2014).

## Material contracts of firms

### Alliance contract databases

The most widely used database for the study of alliances is Securities Data Company (SDC) (Refinitiv)[6]; however, it does not offer full contract texts. Alliance scholars working on contracts tend to use the MERIT-CATI (now discontinued), Clarivate Cortellis Deals Intelligence Analytics

[2] The H4N or Harvard IV-4 Tag is an automatic semantic dictionary used in text mining that classifies words into positive or negative and helps analyze the overall tone of a document.

[3] http://rankandfiled.com/.

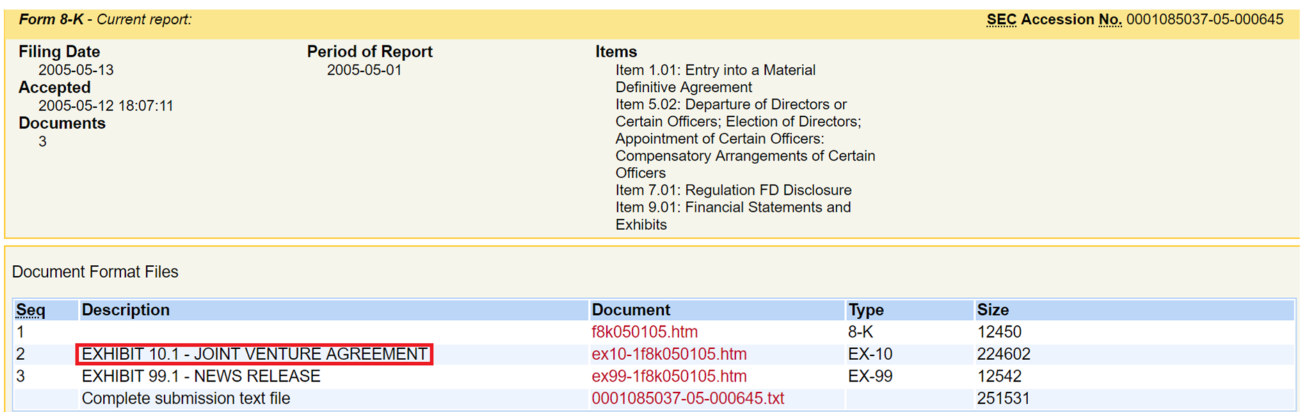
[4] http://www.secinfo.com/.

[5] http://www.secfilings.com/.

[6] https://www.refinitiv.com/en/products/sdc-platinum-financial-securities/.

**Table 1** Alliance contract databases. strengths and weaknesses

| Alliance Database | Strengths | Limitations |
|---|---|---|
| Current Agreements | Full contract texts<br>Good searchability<br>Overview of the alliance | Only healthcare and biopharma alliances<br>Paid subscription |
| Clarivate Cortellis Deals Intelligence Analytics | Full contract texts<br>Good searchability<br>Deal summaries | Only biopharma alliances<br>Paid subscription |
| MERIT-CATI | Historical data from 1960's<br>Agreement summaries<br>Multisector | No text of alliance contracts<br>Focuses on R&D alliances<br>Discontinued |
| Securities Data Company (SDC) | Well organized<br>Excellent searchability<br>Multisector | No text of alliance contracts<br>Paid subscription |
| SEC - EDGAR | Full contract texts<br>Free to access<br>Multisector<br>All alliance types | Not a database per se<br>Information difficult to access and download<br>Possible bias toward publicly traded firms |



*Form 8-K - Current report:* **SEC Accession No.** 0001085037-05-000645

**Filing Date**
2005-05-13
**Accepted**
2005-05-12 18:07:11
**Documents**
3

**Period of Report**
2005-05-01

**Items**
Item 1.01: Entry into a Material Definitive Agreement
Item 5.02: Departure of Directors or Certain Officers; Election of Directors; Appointment of Certain Officers: Compensatory Arrangements of Certain Officers
Item 7.01: Regulation FD Disclosure
Item 9.01: Financial Statements and Exhibits

Document Format Files

| Seq | Description | Document | Type | Size |
|---|---|---|---|---|
| 1 | | f8k050105.htm | 8-K | 12450 |
| 2 | EXHIBIT 10.1 - JOINT VENTURE AGREEMENT | ex10-1f8k050105.htm | EX-10 | 224602 |
| 3 | EXHIBIT 99.1 - NEWS RELEASE | ex99-1f8k050105.htm | EX-99 | 12542 |
| | Complete submission text file | 0001085037-05-000645.txt | | 251531 |

**Fig. 1** 8-K filing details with accurate description of exhibits. Joint venture agreement. *Source*: https://www.sec.gov/Archives/edgar/data/1075636/000108503705000645/0001085037-05-000645-index.htm (accessed: 08/18/2019)

(formerly Thomson Reuters Recap),[7] and Current Agreements[8] databases (Choi and Contractor 2016; Delerue and Sicotte 2018; Kwon et al. 2016). Each of these has advantages and limitations as summarized in Table 1. Furthermore, most of their information is actually extracted from EDGAR and few researchers and practitioners are aware that this information is public. For more details on alliance databases, refer to Schilling (2009). Since the EDGAR electronic filing and retrieval system was implemented, researchers and investors dramatically increased their use of data made available by the SEC. For example, legal and IT scholars began extracting and analyzing material contracts (Chen and Bharadwaj 2009; Sanga 2014), and, in the strategy field, Hegde (2014) used an algorithm to extract and analyze license contracts. In line with the latter, we used an algorithm to extract JV contracts from the SEC. However, in contrast to the previous work, including that of Loughran and McDonald (2016), our work is based on the use of text phrases or keywords, with the purpose of not only analyzing the text, but also retrieving it from Edgar, as opposed to using a predetermined list of words to perform sentiment analysis in already available texts.

## Contract disclosure and the search for JV agreements

US publicly traded firms are required to file their annual reports with the SEC using form 10-K (20-F for foreign firms), their quarterly reports through form 10-Q and their

[7] https://clarivate.com/cortellis/solutions/deals-intelligence-analytics/.

[8] https://www.currentagreements.com/.

*Form 10QSB* - *Optional form for quarterly and transition reports of small business issuers:* **SEC Accession No.** 0001144204-06-022093

**Filing Date**
2006-05-22
**Accepted**
2006-05-22 16:52:09
**Documents**
14

**Period of Report**
2006-03-31

Document Format Files

| Seq | Description | Document | Type | Size |
|---|---|---|---|---|
| 1 | | v044058_10qsb.txt | 10QSB | 132990 |
| 2 | | v044058_ex3-1.htm | EX-3.1 | 774 |
| 3 | GRAPHIC | v044058_ex3-1x1x1.jpg | GRAPHIC | 93199 |
| 4 | GRAPHIC | v044058_ex3-1x2x1.jpg | GRAPHIC | 68740 |
| 5 | GRAPHIC | v044058_ex3-1x3x1.jpg | GRAPHIC | 91701 |
| 6 | GRAPHIC | v044058_ex3-1x4x1.jpg | GRAPHIC | 84909 |
| 7 | | v044058_ex3-2.txt | EX-3.2 | 32016 |
| 8 | | v044058_ex4-1.txt | EX-4.1 | 2589 |
| 9 | | v044058_ex4-2.txt | EX-4.2 | 2461 |
| 10 | | v044058_ex10-2.txt | EX-10.2 | 3419 |
| 11 | | v044058_ex31-1.txt | EX-31.1 | 3400 |
| 12 | | v044058_ex31-2.txt | EX-31.2 | 3427 |
| 13 | | v044058_ex32-1.txt | EX-32.1 | 1583 |
| 14 | | v044058_ex32-2.txt | EX-32.2 | 1597 |
| | Complete submission text file | 0001144204-06-022093.txt | | 653099 |

**Fig. 2** 10QSB filing details description of exhibits left blank. *Source*: https://www.sec.gov/Archives/edgar/data/923771/000114420406022093/0001144204-06-022093-index.htm (accessed: 08/18/2019)

earning releases with form 8-K. In addition, companies must disclose material contracts, such as management, license and JV agreements (Bommarito et al. 2018; Chen and Bharadwaj 2009; Hegde 2014; Sanga 2014). Contracts should be attached as additional files to annual or quarterly report filings (10-K, 10-Q) and should be referenced in item 10 of the exhibits section of the report (Overdahl 1991). Therefore, a given filing contains a main document with several additional files attached as seen in Fig. 1. In this case, one can see a filing of an 8-K report in which the main document is attached to sequence 1, followed by two additional documents: a joint venture contract in sequence 2 (exhibit 10.1) and a news release in sequence 3 (exhibit 99.1). Hence, to find joint venture agreements, one must enter a company name in the EDGAR search engine and explore dozens of filings and Sect. 10 of their exhibits, hoping to find these contracts. This manual process is slow and inadequate for the creation of reliable contract databases. Therefore, an automated solution to this problem is the development of a search program in the *PYTHON* programming language.

The program we called *GetEdgar* was written to mimic the steps of a manual company search, but since there was no list of all the thousands of company names, the program was instructed to search by Standard Industrial Classification (SIC) code, an option provided by the advanced search in EDGAR. Therefore, first all 444 SIC numbers were downloaded, and the code was programmed to sequentially open each SIC code. This search shows companies in each industry, wherein each company and its filings can be accessed. The program was then coded to open the URL of each company for each SIC code and each of their 10-K, 10-Q and 8-K filings.

The next step in the code searches for joint ventures in the descriptions of the attachments to the filings (see Fig. 1) by using a set of predetermined keywords such as joint venture contract, JV contract or agreement or a combination of similar terms. When a match is found, the program includes it in an *Excel* file with the URL where the contract is publicly hosted, the link to this URL being in the *document* column of the filing, as in the case of Fig. 1 where it says *ex10-1f8k050105.htm.* Therefore, the program is also instructed to find the match and copy the URL from the *description* column to retrieve the link.

The first automatic search included all foreign and US-based companies required to report to EDGAR (over 670,000) spanning 444 different industry classifications between 2000 and 2016. This search included the most common filings, namely the 10-K, 10-Q, 8-K and 20-F. The results yielded a disappointing 53 JV agreements. Further manual exploration of EDGAR found that companies misfile contracts in other filings such as 10-QSB (Fig. 2), 10 KSB, 6-K and S-4. Hence, the search was expanded to over 30 classes of filings, resulting in 359 JV agreements.

Additionally, it was found that many companies do not properly describe their attached contracts and give them generic names or give no description to the attachments at all, as in Fig. 2 where the contract is attached to *Seq 10* and the description left blank. As such, neither a person nor the algorithm can identify it without opening each of the 14 documents of the filing and exploring their contents.

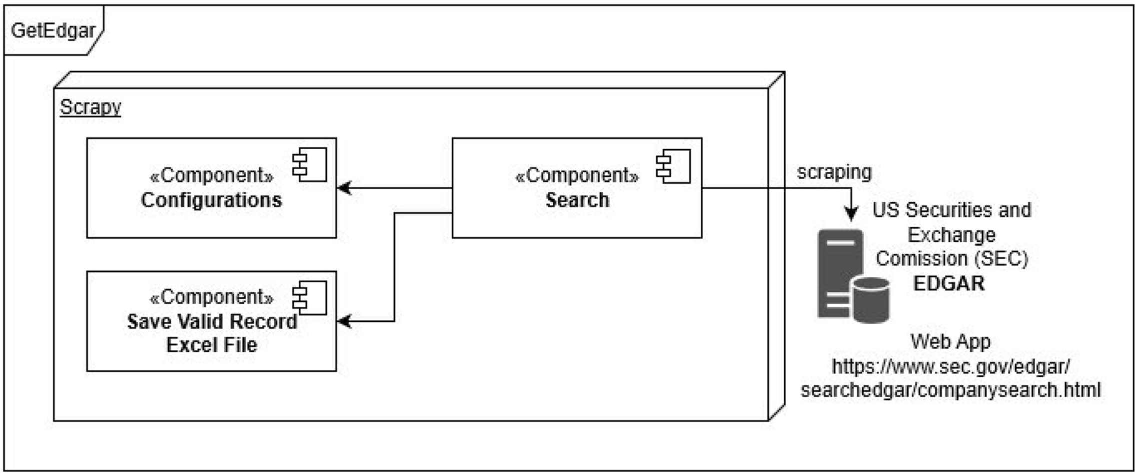



**Fig. 3** Component/context UML diagram

To find these hidden contracts, a final tweak to the program was made. The additional coding included opening of and performing text analysis on each of the attachments to the filings of all companies. The text analysis would search for the same joint venture-related keywords, not in the description of the filing but inside the first 500 hundred words of each document. This method proved successful and provided more contracts but also documents such as letters and press releases with the words joint venture in the headlines. These entries had to then be deleted manually from the Excel worksheet; therefore, we included an extra piece of code with unwanted keywords to reduce the number of such false entries.

## GetEdgar's architecture and artifact framework

We developed the following artifact framework with three unified modeling language (UML) diagrams. Figure 3 is a component diagram that provides an overview of our GetEdgar program. It includes three internal and one external components. The internal components are 1) Search, which is the main component (code). It builds a scraping spider that constructs and searches EDGAR URLs based on configurations for industries (SICS), companies (CIKS), and filing types (Keydocs). 2) Configurations is where search settings are saved in a .json file and allow the application to operate properly. This file contains several editable search criteria, including the list of SICS (industries), filings types and keywords. These are basically the user preferences and search criteria for the entire program. 3) The Save Valid Record component is the program output and stores the contracts found that fulfill the search criteria (Keywords, false keywords, search depth) in an Excel file with an editable file name. Finally, the external component is the EDGAR database itself. Thus, the search component interacts several times with the external component based on user configurations and provides output in an Excel file.

Figure 4 is a class/package diagram that displays a logical view of the project with classes, their methods and their associations. This project is object oriented, and each box in the diagram represents a class. A class is an abstract representation of an object. An object is composed of attributes and functions. There are six classes: ReadJson, GetEdgarItem, GetEdgarSpider, Reading, Save file and log. They are contained in four packages: configurations, src (source)/utils (utilities), spiders, src/response. This diagram helps understand the application's internal logic.

Figure 5 displays a UML sequence diagram that shows the process flow from filter configurations to storage of the information found in the excel output file. After adjusting the search criteria or preferences in configurations (.json file), an actor initiates the program. The program launches a spider and searches EDGAR using both SICs (industries) and all CIKs (company codes). The program verifies the configurations, saves them and begins creating URLs as a regular user would when clicking through filings and attachments. It does this for all SICs and all company codes (CIKs). The print statements show the process while the program is running. The program basically opens all industries, companies and filings, examines their attachments and validates against the search keywords. During the process, if the program finds a contract that matches the search criteria, it stores its information in a file. The circle in the middle represents a loop that repeats until it searches all SICS and CIKS in the database according to the user search criteria. Thus, if the configurations are set to search for all industries (SICs), the program will actually run through the entire EDGAR database.

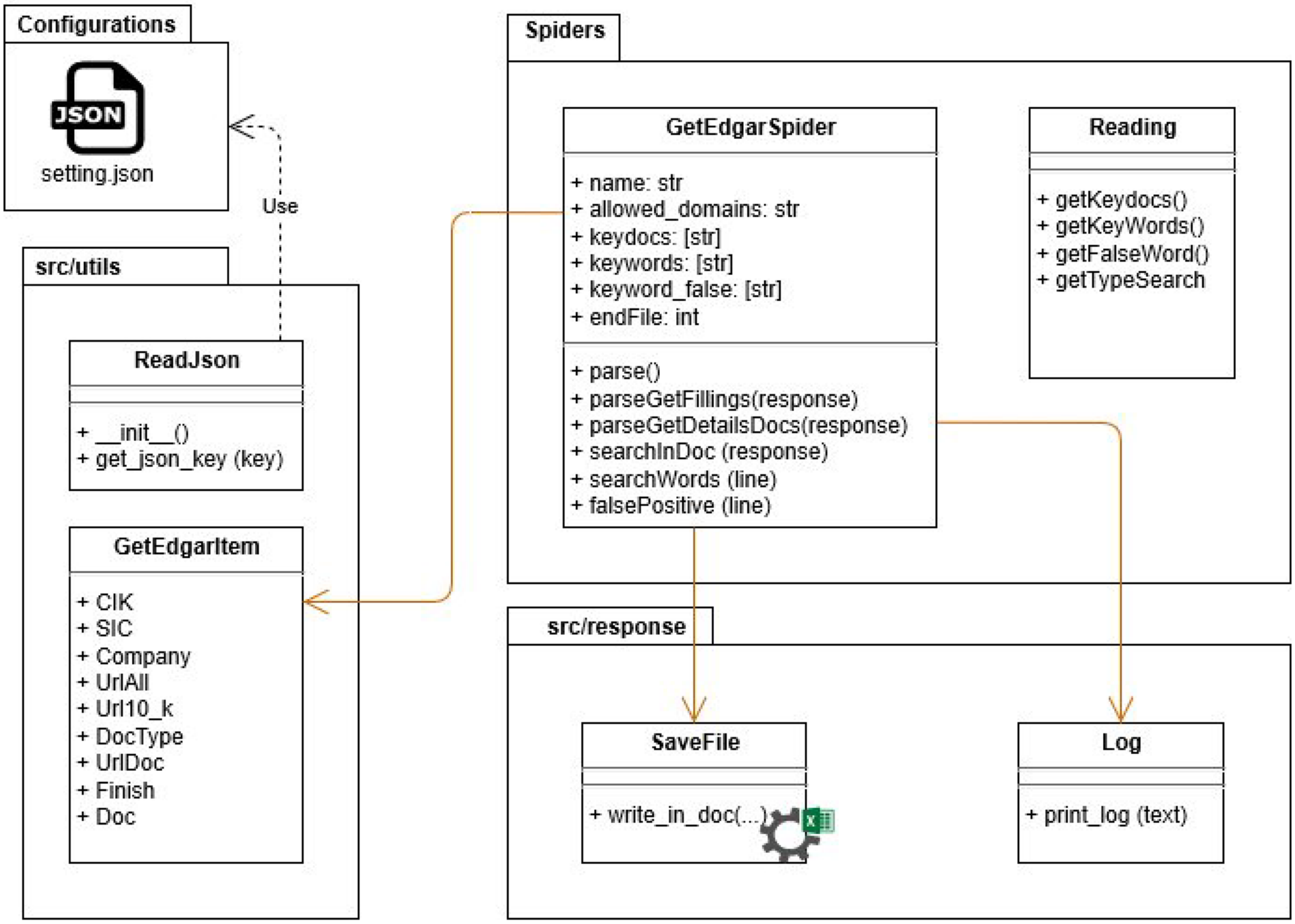



**Fig. 4** Class/package UML diagram

## GetEdgar's location, installation and operation

The GetEdgar project is publicly available and can be accessed and downloaded from https://github.com/getedgar/filingsearch.

To run the program, Python version 2.7 or higher must be installed as well as the Scrapy package. We created a *Manual.html* file that contains more detailed installation information for program operation. Before running the program, the user can modify the search criteria located by accessing the /configurations/ path and opening the setting.json file.

**setting.json file keys**

The search_depth_in_file key indicates the depth of the search to be performed on documents. The greater the depth, the more documents can be found, but the process will be slower. Therefore, the initial part of the contracts where the specific keywords are usually mentioned is analyzed by the program. It has a default depth of 500 words but this can be easily modified:

```
"sics": [
  "100","200","800","900","1000","1040","1090","1220","1221"
]
```

Enter all the SIC numbers to be searched against this key.

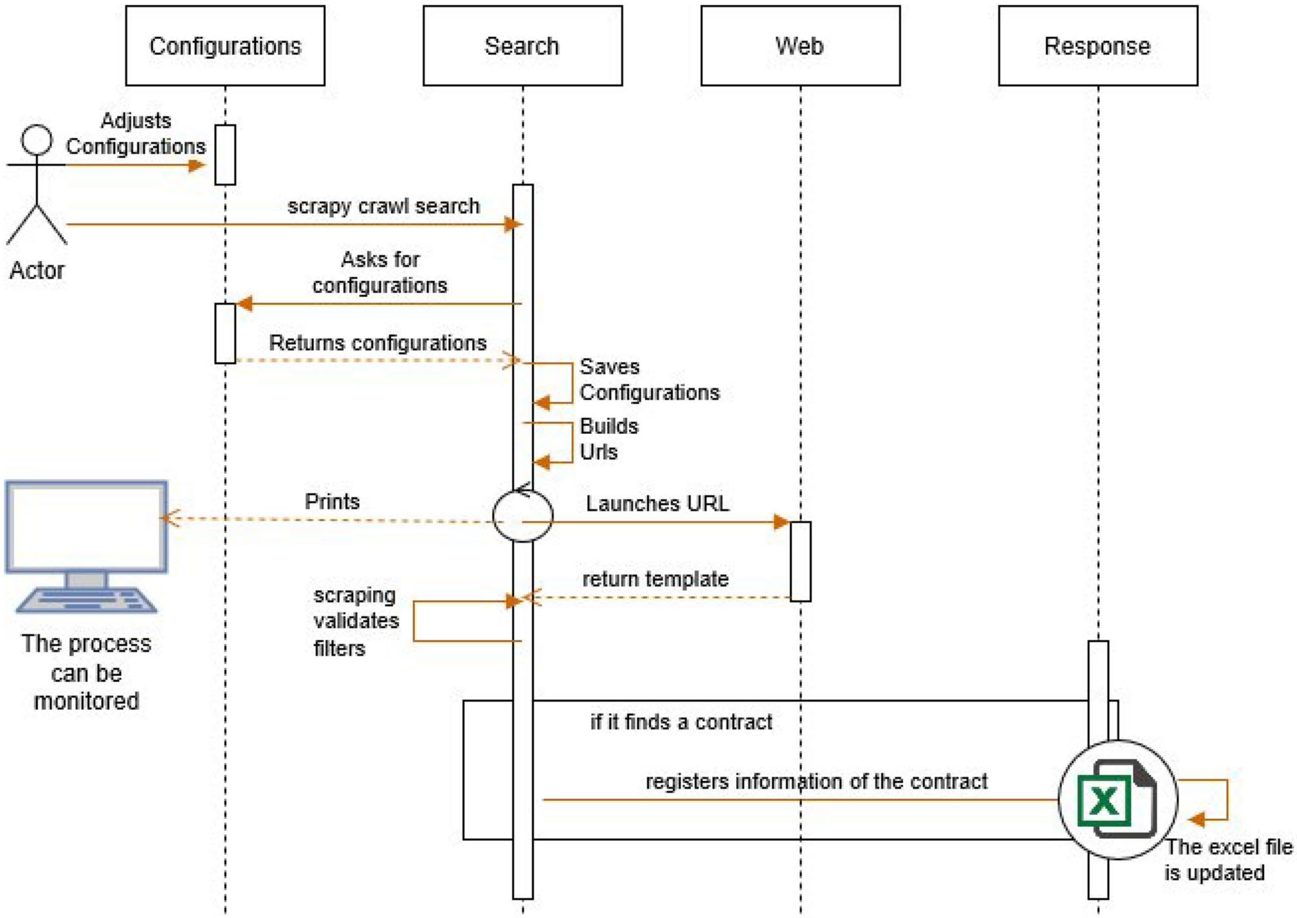
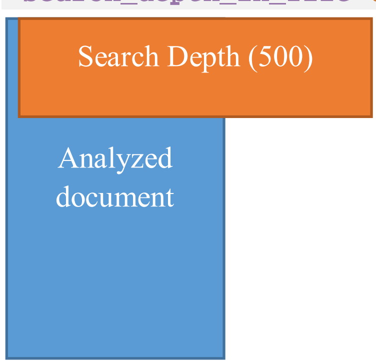



**Fig. 5** Sequence UML diagram

```
"search_depth_in_file": 500,
```




The program can be set to search-specific filings (key_docs) as seen in the code:

```
"key_docs": [
  "10-k","10-Q","8-k","S-8","20-F","F-10","F-8","6-k","S-1/A"
]
```

**Table 2** Variables from the firms' EDGAR filings

| Variable | Description |
|---|---|
| Central index key (CIK) | SEC identification number for the filing company |
| Contract URL | URL where the contract is made public |
| Country | Country where filing company is headquartered |
| Exhibit | Exhibit number where the firm attaches the contract (e.g., 10.3) |
| Filer address | Address of filing company |
| Filing company | Name of filing firm |
| Date of filing | Date the company filed the report to the SEC |
| Filing description | Description of the type of filing the company is reporting |
| Filing URL | URL where the main document and its attachments can be accessed |
| SIC code | Four-digit standard industry classification code of filing company |

The list of search keywords is also editable and is where the user enters the type of document they want to search for.

```
"keywords": [
  "joint venture master agreement",
  "joint venture agreement"
]
```

There is also a false word list which, unlike the keywords, indicates that a document is not useful for our purpose. These are also known as false positives.

```
  "press release"
]
```

If a user wants to perform a specific company search, the "**active**" setting must be set to true in the code. A CIK or a list of the CIK codes corresponding to each company must then be entered.

```
"company_specifies":{
  "active": false,
  "__read": "This obtains the CIK numbers of the companies
https://www.sec.gov/edgar/searchedgar/cik.htm",
  "companies_CIKs":[
    "0000788206","other"
  ]
}
```

The file_out key is used to edit the name of the Excel output file. For example:

```
"file_out": {
  "name": "JVContractsDataBase"
}
```


**Fig. 6** Word cloud of 626 joint venture agreements

Finally, after entering the search criteria, users must open a terminal or a command line prompt, navigate to the /spiders/ project folder which contains the program that will deploy a webcrawler across several webs to perform the scraping, and enter the command ***scrapy crawl search.***

## Dissecting and analyzing the JV contracts

### The final JV contract sample

Following the filtering of the Excel worksheet, the final JV contract database consisted of 626 agreements. The Excel output file created by the program included a series of variables as seen in Table 2.

Other JV information was obtained directly from the agreements. For example, using the two-digit industrial classification code (SIC), JV industries are identified as manufacturing (47%), mining (30%), services (11%), real estate (5%), transportation and communications (4%) and retail (3%). In terms of the JVs' objective, the sample can be classified into production (34%), development (27%), exploitation (28%), distribution (12%) and research and development (14%). Three hundred (47.9%) of the JVs are domestic, while 326 (52.1%) are international. In terms of governance, it is interesting to see that 164 JVs have not set up a board of directors and only in 108 are their partners' equity shares equal to their representation on the board.

Additional variables can be extracted from the contracts and matched to country-level data or institutional variables to perform different types of analysis. These contracts indeed hold a wealth of information, and the variables that can be extracted for strategic analyses or future research studies are enormous. For example, ample possibilities exist for research on governance mechanisms in joint ventures, control, choice of law, boards of directors, safeguards, arbitration, contractual complexity, etc. Additionally, these contracts can serve as templates for practitioners when negotiating and crafting their own joint ventures or any other contract available in the SEC's EDGAR repository.

### Dissecting the joint venture contracts

In text mining, a common first approach to visually represent word frequency of a document or collection of documents is a word cloud. This identifies key terms by displaying word frequency in terms of size (Chi et al. 2015). Figure 6 shows an initial dissection of the joint venture dataset. This graph was generated using *R* statistical software taking into account the term frequency-inversed document frequency (tf-idf) metric, which does not merely show the frequency of words in all contracts but calculates the importance of a term in a contract relative to the entire contract collection. This calculation avoids giving too much importance to a word that is highly repeated in just one or a few documents but not much in others (Robertson 2004). Figure 6 clearly shows relevant and recurring terms in the JV contracts such as *board, parties*, *shares*, *project*, *meeting* and *management*, shining a light on the importance of certain clauses for JV governance.

A second dissection was performed focusing on JV contractual clauses or provisions. The contracts were read, and their clauses were identified and marked. Figure 7 displays the basic 48 clauses present in the contracts and their inclusion percentage. The least used clauses are *hardship* in five contracts (0.008%), *time deadlines* in 31 contracts (4.9%), *put options* in 49 contracts (7.8%), and *change in control of a party* in 75 contracts (12%). The most common clauses in the 626 JV contracts are *description of the parties* in 591 contracts (94%), *applicable law* in 602 contracts (96%), followed by *scope* in 605 contracts (97%) and the most common clause *type of joint venture* which is present in 622 contracts (99%).

As shown in Fig. 7, JV contracts may include a wide array of clauses ranging from type of alliance and scope, as the most common, to deadlines and deadlocks as the least. This dissection of the agreements serves as a foundation to explore their anatomy and permits an initial visualization and starting point for further analyses that will improve an understanding of contract design, JV governance and strategy.

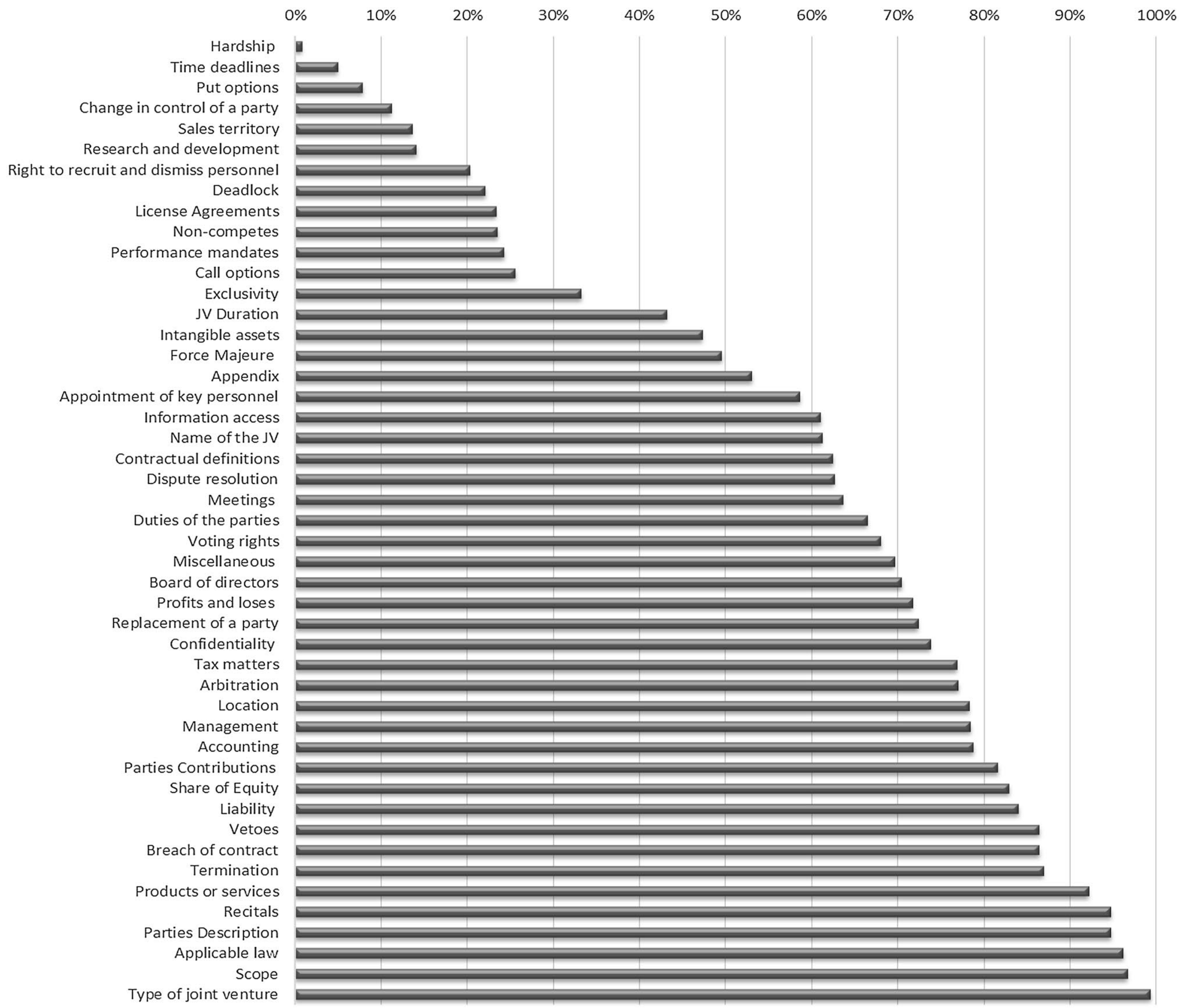



**Fig. 7** Types of clauses present in joint venture agreements

## Discussion

A contract is a legal document that defines the governance structure of most inter-firm alliances, as the framework under which the parties will cooperate (Schepker et al. 2014). Joint ventures, unlike other types of alliances, are governed by detailed contracts describing how the relationship within the alliance will be structured and governed (Reuer et al. 2016). Only recently have researchers begun to take interest in the microlevel details of agreements as a means to understand their structure and interactions in terms of governance mechanisms (Contractor and Reuer 2014). Other aspects that have received attention include information asymmetry, uncertainty and bounded rationality (Hart and Moore 2007). Nevertheless, empirical understanding of contract design and crafting, as well as how they work, remains limited, due to both their complexity and the difficulty of accessing full alliance contracts (Schepker et al. 2014).

Access to firsthand information has been commonly used by investors and management teams for contract design and crafting improved corporate strategies. This article contributes to these aims by describing and exploring a vast repository of contracts, which provide valuable firm-level information via example joint venture contracts using the EDGAR electronic system for online search and retrieval. It has been shown how investors and researchers can use public data to advance knowledge and make investment decisions, particularly since the implementation of the EDGAR online system which allows access to such data in a timely and expedited manner if the user employs the strategies suggested in this article. As demonstrated for the case of joint venture contracts, an improved understanding of how EDGAR works, how the information is stored and its search engine functions allows accessing company data manually or using an automated program.

Considering that the information contained in EDGAR is vast and complex, some researchers have developed tailored programs to access such content (Garcia and Norli 2012a). The purpose of this paper is to encourage researchers to use this source to create new datasets, and to that end, the algorithm used for this study is available upon request. The strategy described herein is just one of many possible ways to systematically search and download information contained in the reports and accompanying documents. The programs can be adapted to access other types of documents and contracts. They are relatively simple to write, requiring only basic programming skills in Java, Python or other programming languages.

## Conclusions and avenues for further research

This paper describes the use of programming and text mining techniques to access public information in the US Securities and Exchange Commission's EDGAR repository. It explains the purpose of the SEC and its EDGAR filing system, how it works and how anyone can access this information. It further explains how companies file their material contracts as attachments to their filings and how to find them. The paper also outlines the development of a program for mining EDGAR, resulting in the extraction of over six hundred JV contracts. Finally, the paper explores basic text mining techniques to illustrate the most important terms and clauses of the JV contract sample and opens up a broad research agenda.

The joint venture dataset described previously was built using over six hundred joint venture contracts publicly available through EDGAR. The findings were generated based on an exploration of actual contractual clauses, among which the most frequently used include the type of joint venture, its scope, the laws that govern the contract and the description of the parties (refer to Fig. 4 for a complete overview). This information may serve several purposes, which at the same time configures avenues for future research. First, one interesting possibility is to use *the choice of law or applicable law* clause to determine the institutional factors affecting contract complexity and enforcement; this can be achieved by finding connections with the literature on legal studies (Sanga 2014) and strategy, as well as on alliance negotiations (Contractor and Ra 2000). These links will allow strategy and international business scholars to deepen their understanding of the implications of the choice of law in JV contracts, its negotiation and the factors influencing such a choice. Second, the use of alliance contracts may bring about an enhanced understanding of firm strategies by focusing on the scope and noncompete clauses as means for drawing the line between cooperation and competition, a common dilemma in JV negotiation and execution. Also, the analysis of other provisions and clauses may provide substantial contributions to the fields of strategy, management and international business.

The analysis of the JV contracts addressed here provided a clearer understanding of joint venture governance. The board of directors appeared to be the most commonly used governance mechanism. Nonetheless, we found that many JVs did not include this topic within their contracts. Future research may address the underlying reasons for such an omission and the governance mechanisms that replace such boards.

Finally, it is important to mention that many contracts were neither filed or posted using the appropriate headings, making them very difficult to find. It would be fascinating to investigate why many firms misfile, intentionally or not, their material contracts and consider the implications of this behavior for the literature on accounting disclosures and strategy. Finally, it would seem that, although the SEC closely monitors the submission of annual and quarterly reports, further controls are needed when it comes to exhibits and documents.

**Acknowledgements** The authors would like to thank the reviewers and editors of IJDG for their time and constructive feedback; engineers Juan David Arboleda and Juan Felipe Muñoz for the programming support; and Tjebbe Donner for proofreading the final version of the manuscript.

## Compliance with ethical standards

**Conflict of interest** On behalf of all authors, each corresponding author states that there is no conflict of interest.

## References

Argyres, N.S., J. Bercovitz, and K.J. Mayer. 2007. Complementarity and evolution of contractual provisions: An empirical study of IT services contracts. *Organization Science* 18(1): 3–19.
Bommarito, M.J., D.M. Katz, and Detterman, E.M. 2018. *OpenEDGAR: Open source software for SEC EDGAR analysis*. Available at SSRN 3194754.
Bushee, B.J., and C. Leuz. 2005. Economic consequences of SEC disclosure regulation: Evidence from the OTC bulletin board. *Journal of Accounting and Economics* 39(2): 233–264.
Chen, Y., and A. Bharadwaj. 2009. An empirical analysis of contract structures in IT outsourcing. *Information Systems Research* 20(4): 484–506.
Chi, M., S. Lin, S. Chen, C. Lin, and T. Lee. 2015. Morphable word clouds for time-varying text data visualization. *IEEE Transactions on Visualization and Computer Graphics* 21(12): 1415–1426.
Choi, J., and F.J. Contractor. 2016. Choosing an appropriate alliance governance mode: The role of institutional, cultural and

geographical distance in international research & development (R&D) collaborations. *Journal of International Business Studies* 47(2): 210–232.
Cong, Y., A. Kogan, and M.A. Vasarhelyi. 2007. Extraction of structure and content from the edgar database: A template-based approach. *Journal of Emerging Technologies in Accounting* 4(1): 69–86.
Contractor, F.J., and W. Ra. 2000. Negotiating alliance contracts: Strategy and behavioral effects of alternative compensation arrangements. *International Business Review* 9(3): 271–299.
Contractor, F.J., and J.J. Reuer. 2014. Structuring and governing alliances: New directions for research. *Global Strategy Journal* 4(4): 241–256.
Dalton, D., S. Buchheit, D. Oler, and M. Zhou. 2013. Enforcement mechanisms for SEC reporting deadlines. *Research in Accounting Regulation* 25(2): 185–195.
De Franco, G., M.F. Wong, and Y. Zhou. 2011. Accounting adjustments and the valuation of financial statement note information in 10-K filings. *The Accounting Review* 86(5): 1577–1604.
Delerue, H., and H. Sicotte. 2018. Formal international contracts in the presence of cultural distance: An empirical analysis of biopharmaceutical alliances. *Thunderbird International Business Review* 61(4): 595–607.
Easton, P.D., and M.E. Zmijewski. 1993. SEC form 10K/10Q reports and annual reports to shareholders: Reporting lags and squared market model prediction errors. *Journal of Accounting Research* 31: 113–129.
Garcia, D., and Ø. Norli. 2012a. Crawling EDGAR. *The Spanish Review of Financial Economics* 10(1): 1–10.
Garcia, D., and Ø. Norli. 2012b. Geographic dispersion and stock returns. *Journal of Financial Economics* 106(3): 547–565.
Gerdes, J. 2003. EDGAR-analyzer: Automating the analysis of corporate data contained in the SEC's EDGAR database. *Decision Support Systems* 35(1): 7–29.
Gong, Y., O. Shenkar, Y. Luo, and M. Nyaw. 2007. Do multiple parents help or hinder international joint venture performance? The mediating roles of contract completeness and partner cooperation. *Strategic Management Journal* 28: 1021–1034.
Griffin, P.A. 2003. Got information? investor response to form 10-K and form 10-Q EDGAR filings. *Review of Accounting Studies* 8(4): 433–460.
Hart, O., and J. Moore. 2007. Incomplete contracts and ownership: Some new thoughts. *American Economic Review* 97(2): 182–186.
Hegde, D. 2014. Tacit knowledge and the structure of license contracts: Evidence from the biomedical industry. *Journal of Economics & Management Strategy* 23(3): 568–600.
Kwon, S., J. Haleblian, and J. Hagedoorn. 2016. In country we trust? national trust and the governance of international R&D alliances. *Journal of International Business Studies* 47(7): 807–829.
Loughran, T., and B. McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-Ks. *The Journal of Finance* 66(1): 35–65.
Loughran, T., and B. McDonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research* 54(4): 1187–1230.
Loughran, T., and B. McDonald. 2017. The use of EDGAR filings by investors. *Journal of Behavioral Finance* 18(2): 231–248.
Luo, Y. 2005. Transactional characteristics, institutional environment and joint venture contracts. *Journal of International Business Studies* 36(2): 209–230.
Overdahl, J.A. 1991. A researcher's guide to the contracts of firms filing with the SEC. *The Journal of Law & Economics* 34(2): 695–701.
Reuer, J.J., A. Ariño, L. Poppo, and T. Zenger. 2016. Alliance governance. *Strategic Management Journal* 37(13): E37–E44. https://doi.org/10.1002/smj.2535.
Robertson, S. 2004. Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation* 60(5): 503–520.
Rogers, J.L., D.J. Skinner, and S.L. Zechman. 2017. Run EDGAR run: SEC dissemination in a high-frequency world. *Journal of Accounting Research* 55(2): 459–505.
Sampson, R.C. 2004. The cost of misaligned governance in R&D alliances. *Journal of Law Economics and Organization* 20(2): 484–526.
Sanga, S. 2014. Choice of law: An empirical analysis. *Journal of Empirical Legal Studies* 11(4): 894–928.
Schepker, D.J., W. Oh, A. Martynov, and L. Poppo. 2014. The many futures of contracts: Moving beyond structure and safeguarding to coordination and adaptation. *Journal of Management* 40(1): 193–225.
Schilling, M.A. 2009. Understanding the alliance data. *Strategic Management Journal* 30(3): 233–260.
You, H., and X. Zhang. 2009. Financial reporting complexity and investor underreaction to 10-K information. *Review of Accounting Studies* 14(4): 559–586.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.