Understanding Financial Information Seeking Behavior from User Interactions with Company Filings

Ariannezhad, M.; Yahya, M.; Meij, E.; Schelter, S.; de Rijke, M.

DOI
10.1145/3487553.3524636

Publication date
2022

Document Version
Author accepted manuscript

Published in
WWW '22 Companion

Understanding Financial Information Seeking Behavior from User Interactions with Company Filings

Mozhdeh Ariannezhad*
m.ariannezhad@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Mohamed Yahya
myahya6@bloomberg.net
Bloomberg
London, United Kingdom

Edgar Meij
emeij@bloomberg.net
Bloomberg
London, United Kingdom

Sebastian Schelter
s.schelter@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Maarten de Rijke
m.derijke@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

ABSTRACT

Publicly-traded companies are required to regularly file financial statements and disclosures. Analysts, investors, and regulators leverage these filings to support decision making, with high financial and legal stakes. Despite their ubiquity in finance, little is known about the information seeking behavior of users accessing such filings. In this work, we present the first study of this behavior. We analyze 14 years of logs of users accessing company filings of more than 600K distinct companies on the U.S. Securities and Exchange Commission's (SEC) Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system, the primary resource for accessing company filings. We provide an analysis of the information-seeking behavior for this high-impact domain. We find that little behavioral history is available for the majority of users, while frequent users have rich histories. Most sessions focus on filings belonging to a small number of companies, and individual users are interested in a limited number of companies. Out of all sessions, 66% contain filings from one or two companies and 50% of frequent users are interested in six companies or less. Understanding user interactions with EDGAR can suggest ways to enhance the user journey in browsing filings, e.g., via filing recommendation. Our work provides a stepping stone for the academic community to tackle retrieval and recommendation tasks for the finance domain.

ACM Reference Format:
Mozhdeh Ariannezhad, Mohamed Yahya, Edgar Meij, Sebastian Schelter, and Maarten de Rijke. 2022. Understanding Financial Information Seeking Behavior from User Interactions with Company Filings. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3487553.3524636

*Research conducted when the author was doing an internship at Bloomberg.

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9130-6/22/04…$15.00
https://doi.org/10.1145/3487553.3524636

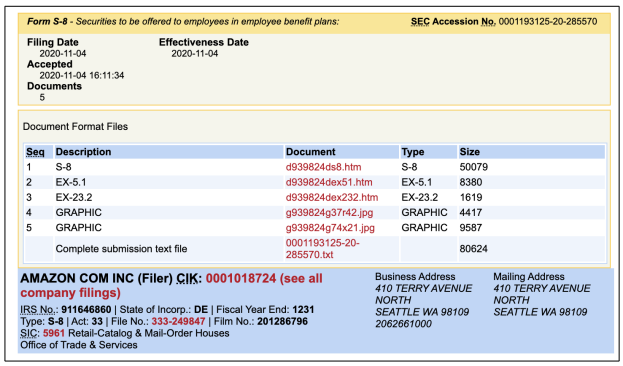
1 INTRODUCTION

Finance is a domain characterized by a large number of financial and regulatory documents, which analysts and investors need to consume as part of their regular workflows. Advanced techniques for search, filtering, and recommendation are crucial to ensure timely access to the right information [3, 13]. Despite this, information retrieval (IR) research focused on the financial domain is still in its early days. While there are some studies on stock recommendation [6, 7, 34], financial entity extraction [24], financial event representation learning [8], ranking [12], and prediction [33], we understand relatively little about how users interact with financial information systems. In this work we take a first step in this direction by focusing on company filings.

Company filings are financial statements of companies or disclosures made by parties tied to these companies. They are a primary source for investors, analysts, advisors, and regulators to acquire information about a company. In the US, all public companies are required to submit filings to the Securities and Exchange Commission (SEC). These filings are exposed to users through the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) database [18]. With an average of over 3,000 filings being submitted per day and hundreds of thousands of daily filing views by users, EDGAR plays a central role in the collection and distribution of financial information.

The Securities and Exchange Commission (SEC) made the EDGAR Log File Dataset (EDGAR-LFD) publicly available. The dataset records access to company filings in the period between February 14, 2003 and June 30, 2017.[1] It captures access to individual filings from different users, alongside meta-information about the filing that is being accessed. The availability of EDGAR-LFD has led to numerous studies in the finance literature [9, 10, 19, 20, 23, 29], which shows the importance of this dataset for research in finance. The focus in such publications is on revealing correlations between the information acquisition of EDGAR users at an aggregate level and financial variables such as stock returns [9, 19, 29]. Different from previous work in the finance literature, we aim to understand how users interact with EDGAR with the focus on information access itself.

[1] https://www.sec.gov/dera/data/edgar-log-file-data-set.html



Form S-8 - Securities to be offered to employees in employee benefit plans: **SEC Accession No.** 0001193125-20-285570

Filing Date 2020-11-04 **Effectiveness Date** 2020-11-04
Accepted 2020-11-04 16:11:34
Documents 5

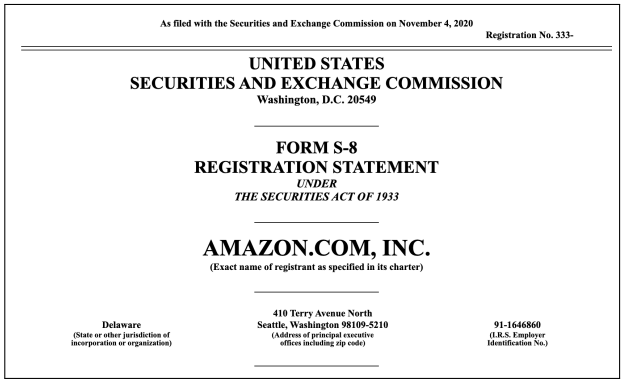
Document Format Files

Seq	Description	Document	Type	Size
1	S-8	d939824ds8.htm	S-8	50079
2	EX-5.1	d939824dex51.htm	EX-5.1	8380
3	EX-23.2	d939824dex232.htm	EX-23.2	1619
4	GRAPHIC	g939824g37r42.jpg	GRAPHIC	4417
5	GRAPHIC	g939824g74x21.jpg	GRAPHIC	9587
	Complete submission text file	0001193125-20-285570.txt		80624

AMAZON COM INC (Filer) CIK: 0001018724 (see all company filings)
IRS No.: **911646860** | State of Incorp.: **DE** | Fiscal Year End: **1231**
Type: **S-8** | Act: **33** | File No.: **333-249847** | Film No.: **201286796**
SIC: **5961** Retail-Catalog & Mail-Order Houses
Office of Trade & Services

Business Address: *410 TERRY AVENUE NORTH SEATTLE WA 98109 2062661000*
Mailing Address: *410 TERRY AVENUE NORTH SEATTLE WA 98109*

(a) Filing index page linking to filing and supplementary files.



As filed with the Securities and Exchange Commission on November 4, 2020
Registration No. 333-

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM S-8
REGISTRATION STATEMENT
UNDER
THE SECURITIES ACT OF 1933

AMAZON.COM, INC.
(Exact name of registrant as specified in its charter)

Delaware	410 Terry Avenue North Seattle, Washington 98109-5210	91-1646860
(State or other jurisdiction of incorporation or organization)	(Address of principal executive offices including zip code)	(I.R.S. Employer Identification No.)

(b) Content of an S-8 filing (corresponds to the first row in Fig. 1a).

Figure 1: Example EDGAR filing index page and content.

In this paper, we provide a comprehensive picture of EDGAR-LFD (Section 3) and analyse how users interact with financial company filings (Section 4). We study the interactions of users with EDGAR on a session level, on a user level, and from a temporal perspective. We find that more than half of the users are one-timers with a single session, while the 10% most active users account for 75% of all sessions. Further, most sessions are focused on filings from a small number of companies. Specifically, 66% of all sessions contain filings from 1–2 companies, suggesting that users tend to focus on filings of a single or of a pair of companies in a session. In the top 10% most frequent users, 50% are interested in six companies or less and 90% of them are interested in 37 or less unique companies over their whole life-cycle, defined as the time between their first and last session on EDGAR. This shows that the most frequent users of the system only focus on a small portfolio of companies.

Our user behavior analysis serves as a stepping stone for the community to tackle retrieval and recommendation tasks for the high-impact financial domain. Our findings have the potential to help financial information providers such as the SEC and commercial providers to better understand the user journey in browsing filings, and suggest ways to enhance the user experience, e.g., via filing recommendation. As a concrete use case to benefit from our analysis, we identify two variations of the filing recommendation task that correspond to the different usage patterns observed in EDGAR, namely next-filing and next-session recommendation.

In summary, we provide the following contributions:

- We provide a detailed description and statistics for EDGAR-LFD, which will inform anyone interested in exploring this dataset (Section 3).
- We provide the first analysis on the information seeking behavior of EDGAR users. Our analysis reveals that user sessions focus on filings from a small number of companies, and that individual users are typically only concerned with a small fraction of the set of all companies (Section 4).
- We discuss the implications of our findings and identify two variations of the filing recommendation task that correspond to the different usage patterns observed in EDGAR, namely next-filing and next-session recommendation (Section 5).

2 RELATED WORK

Previous work on EDGAR-LFD. While this dataset has not been studied by the IR community, there is a considerable amount of work on it in the finance literature. Loughran and McDonald [23] study the consumption of financial information in filings by analyzing the distribution of daily filing requests. Activity on EDGAR is correlated with poor stock performance [10], reactions of the stock market to earnings announcements [20], and predictive of firm performance [9] and stock returns [19, 29]. Co-searches of companies by the same users on EDGAR are used to identify economically related peer firms [18]. These studies examine the usage of EDGAR at an aggregate level, and do not look into user level activities; the focus is on financial variables such as stock returns, and correlations with market events. Unlike previous work, in this paper we analyze the EDGAR-LFD from an information access perspective, in order to understand user behavior and enhance the performance of filing recommendation systems.

Analyzing information seeking behavior. There is a large volume of work on analyzing and learning from interaction logs. Interaction logs are studied to characterize information seeking behavior in different settings such as web search [21], mobile search [31], email search [2], library search [17] and search in productivity software suites [4]. What we add to the work listed above is a comprehensive picture of the information seeking behavior in the finance domain.

Information retrieval in finance. IR in finance has gained attention in the recent years. The FinIR workshop [13] introduces and explores challenges and potential research directions in this area. The FinWeb workshop [2] further explores the usefulness of information on the Web for financial technology. Plachouras et al. [26] and Liu et al. [22] propose search systems specifically designed for financial data. Other related work includes methods to rank financial tweets [5], entity extraction and disambiguation in finance [15], extracting summaries from annual financial reports [1], risk ranking from financial reports [28], and financial document classification [11]. Complementary to existing work, we focus on

[2] https://sites.google.com/nlg.csie.ntu.edu.tw/finweb2021

Table 1: Descriptions of the most frequently accessed filing (form) types on EDGAR.

Type	Description
10-K	Report on a company's performance over a financial year
4	Statement of changes in beneficial ownership of a company
8-K	Report of unscheduled material events or corporate changes at a company
10-Q	Report on a company's performance over a financial quarter

understanding user behavior in interaction with a financial information system and enhancing the user experience with recommender systems.

3 EDGAR & THE LOG FILE DATASET

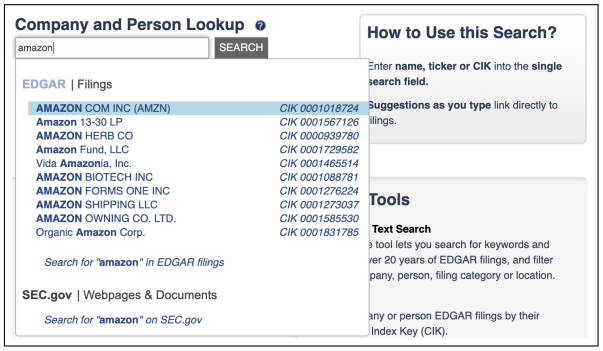
We start by describing the EDGAR system and EDGAR-LFD, which are the main focus of this paper. At the center of EDGAR are *filings*, financial statements of companies and disclosures made by parties tied to these companies. Fig. 1a shows the index page for a specific type S-8 filing issued by Amazon. An S-8 filing is made by companies when they issue equity to their employees. Other examples of filing (form) types are shown in Table 1.[3] A company, like Amazon in our example, is identified by a *unique central index key (CIK)*. The concrete filing is uniquely identified by an *accession number*. As shown in Fig. 1a, the filing itself is composed of multiple files; these are typically the filing document itself in various formats, and supplementary material such as graphics and relevant correspondences. Fig. 1b shows the top of a filing document itself, which corresponds to what a user sees after clicking on a document in the first row of Fig. 1a.

The EDGAR system provides various interfaces for accessing filings. These interfaces can be divided into the following categories:

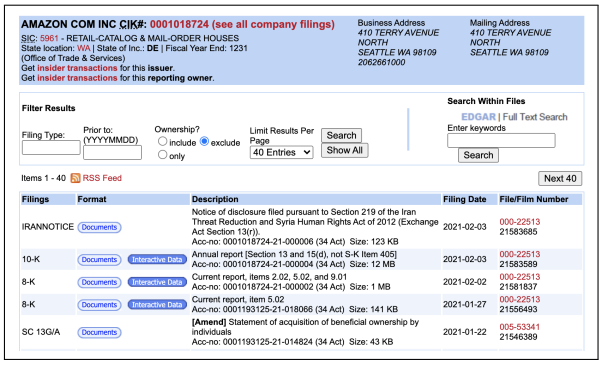
(1) Company lookup (Fig. 2a), to list all filings of that specific company, as shown in Fig. 2b.
(2) Latest filings, for listing filings made in the past few days (Fig. 2c).
(3) EDGAR archive, when the user knows the CIK (and possibly the accession number) of the filing. [4]

These interfaces for reaching a specific filing are reflected in the EDGAR-LFD entries for a specific filing.

With the EDGAR Log File Dataset (EDGAR-LFD), the SEC made access logs for EDGAR filings publicly available.[5] The dataset captures access to individual filings between February 14, 2003 and June 30, 2017, where each record in the data corresponds to a single access by a user to a single filing. A single access corresponds to a single user viewing the contents of a filing (Fig. 1b), or a filing index page (Fig. 1a). Each record contains the date and time of the access and the obfuscated IP address of the user accessing the filing, as well as other details such as the company's CIK and the filing's accession number. An example of an EDGAR-LFD record is shown in Table 2. Additional information about the filings, such as the filing type and the date that the filing was submitted to EDGAR can

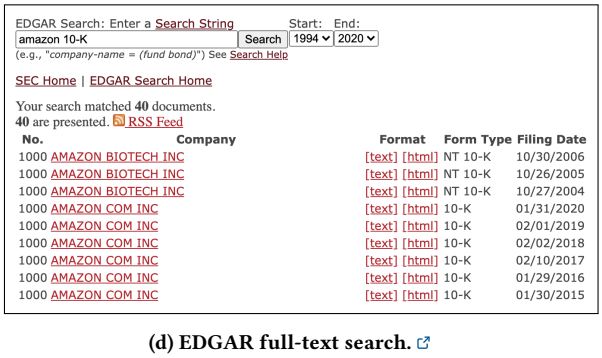


(a) Users can search EDGAR for specific companies.



(b) Chronological list of filings for a company.

(c) Latest EDGAR submissions, chronologically ordered.



(d) EDGAR full-text search.

Figure 2: Examples of specific parts of the SEC's EDGAR website.

[3] Descriptions of different filing types can be found under https://www.sec.gov/forms
[4] https://www.sec.gov/Archives/edgar/data/1018724
[5] https://www.sec.gov/dera/data/edgar-log-file-data-set.html

Table 2: Example of an access record in EDGAR-LFD.

IP	Date	Time	CIK	Accession	Find	Crawler	Extension
203.24.7.aba	2009-10-01	12:09:41	001018724	0001193125-09-154174	3	0	.htm

Table 3: Yearly statistics of EDGAR-LFD after preprocessing.

Year	#valid days	#sessions	#accesses	#unique filings	#unique companies
2003	303	4,421,280	12,377,278	1,430,943	142,960
2004	366	7,377,624	24,697,866	2,012,168	194,072
2005	243	4,505,386	14,290,505	1,817,180	178,416
2006	235	3,971,556	15,424,208	2,232,270	193,222
2007	365	6,703,874	28,643,702	3,467,975	265,258
2008	366	8,351,487	34,072,032	3,469,548	243,016
2009	365	12,111,097	45,109,128	3,244,679	227,435
2010	365	13,633,080	51,405,694	3,403,496	235,163
2011	365	15,543,090	60,586,259	3,598,266	256,870
2012	344	15,835,972	60,817,982	4,745,665	287,858
2013	365	25,724,152	112,685,058	8,517,826	398,015
2014	365	31,280,275	119,560,348	8,158,452	385,115
2015	365	33,084,153	108,967,228	5,993,998	336,679
2016	366	45,364,178	142,694,568	8,211,940	448,807
2017	181	20,282,374	95,057,605	8,586,242	480,512
Total	4,959	248,189,578	926,389,461	11,596,247	607,426

be inferred using the accession number.[6] Appendix A gives more details about the fields of an EDGAR-LFD entry.

We preprocess the dataset and group individual accesses into sessions. The preprocessing and sessionization steps are described in Appendix A. The yearly dataset statistics after the preprocessing steps are shown in Table 3. The processed dataset contains more than 926M accesses to more than 11M filings from 600K companies, grouped into more than 248M sessions, which span across more than 14 years of history.

4 USER BEHAVIOR ANALYSIS

We study financial information seeking behavior by analyzing EDGAR-LFD. We look into temporal access patterns, session-level and user-level behavior. We use the data from all years (bottom row, Table 3). We aim to understand the user behavior and gain insights that can help to improve the user experience in accessing the filings.

4.1 User-level analysis

We focus on user-specific aspects of the filing views. Fig. 3 (Left) shows the cumulative distribution of the number of sessions per user. More than half of the users in the data are one-timers with a single session. The 90th percentile for the number of sessions per user is equal to four, and out of 50.2M users in the data, 5.8M have at least four sessions. We will refer to these top 10% users as *frequent users* in the rest of our analysis. Fig. 3 (Right) displays the cumulative distribution of sessions across users, showing that the 10% most active users account for roughly 75% of the sessions.

[6]From: https://www.sec.gov/Archives/edgar/full-index/

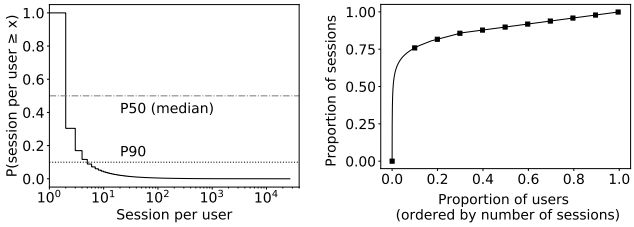


Figure 3: Cumulative distribution of (Left:) number of sessions per user, and (Right:) sessions across users.

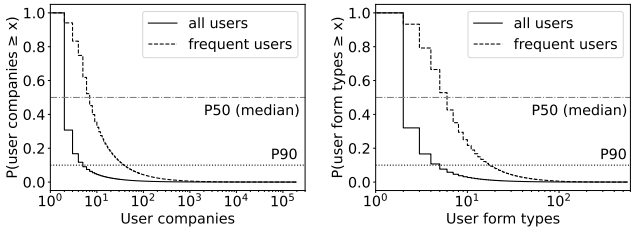


Figure 4: Cumulative distribution of (Left:) number of unique companies, and (Right:) number of unique form types for all users and frequent users.

The analysis shows that the proportion of infrequent users is considerable. For such users, historical behavior beyond the current session is scarce. For downstream tasks such as recommendation, this implies that instant recommendation scenarios such as session-based recommendation, where only the information from the current session is considered, are suitable for a general EDGAR user. We also notice that frequent users are responsible for over 75% of all the sessions in the data, showing that for such users we can rely on rich historical behavior for downstream tasks, and other recommendation scenarios that rely on historical behavior (such as sequential recommendation) could be considered.

We further look into the companies and form types of interest to all users and to the subset of frequent users. Fig. 4 shows the cumulative distribution of unique number of companies and form types that users have interacted with over their whole life cycle. Of the frequent users, 50% are interested in six or less companies and 90% of them are interested in 37 or less unique companies in total; frequent EDGAR users are concerned with only a small subset of companies that have their filings available on the website. The median and 90th percentile are much less if we consider all users. A similar observation holds for the number of unique form types. Out of 505 different form types available, 50% of the frequent users are interested in five or less and 90% are interested in at most 17 unique form types.

This means that each user is concerned with a very small fraction of the data that is available on EDGAR. This implies that in downstream tasks for improving user journey on EDGAR, such as recommendation, focusing on companies and filing types that a

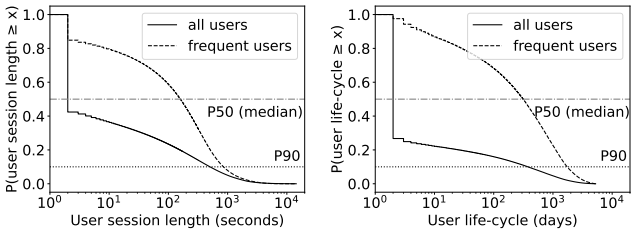


Figure 5: Cumulative distribution of (Left:) average session length per user, and (Right:) life-cycle of users in the whole time frame of the data, for all users and frequent users.

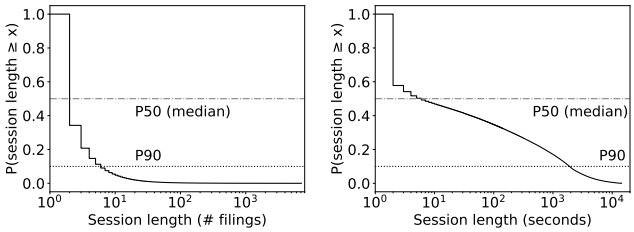


Figure 6: Cumulative distribution of (Left:) session length in terms of number of filings accessed in a session, and (Right:) session length measured in seconds.

user has previously shown interest in will be effective. Such a focus can drastically reduce the search space in different retrieval and recommendation scenarios.

We study the average session length, defined as time between the first and last access in a session, for all users and for frequent users. Fig. 5 (Left) shows the cumulative distribution. The session length is longer for frequent users, with 9 and 15 minutes for the 90th percentile of all and frequent users, respectively.

Our analysis reveals that most of the sessions on EDGAR are short; users tend to focus on a single task in a session, that can translates to accessing a particular filing of interest. There is a notable difference between the session length distribution of frequent users and all users; frequent users have longer sessions, which could indicate that they are a specific group of users, for example analysts. It is worth mentioning that EDGAR-LFD does not provide any information about the type of users who are accessing the filings, but the usages patterns could help to identify different user groups [9].

Fig. 5 (Right) shows the cumulative distribution of life-cycle of users, defined as the number of days between the first and the last session. While the data spans 14 years, the frequent users have a median life-cycle of roughly a year and when considering all users, the 90th percentile for life-cycle is around a year. We observe that frequent users have a much longer life-cycle. For such users, in cases where the historical information is used for a downstream task, the change in the interests of a user in time needs to be accounted for; a very old historical behavior could be less relevant than a recent one when predicting the future interactions of a user.

4.2 Session-level analysis

We analyze the characteristics of sessions on EDGAR in this section. Fig. 6 shows the cumulative distribution of session length in

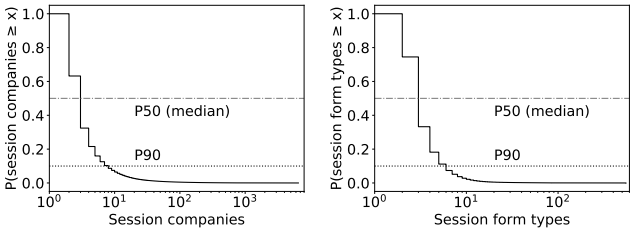


Figure 7: Cumulative distribution of (Left:) number of unique companies, and (Right:) number of unique form types in sessions with more than one filing accessed.

terms of number of filings viewed and session length measured in seconds between the first and the last access in the session. We observe that 68% of all sessions contain only one filing, suggesting that often times users are interested in one filing at a time. This means that most sessions are focused on a single information need, corresponding to a small number of filings at a time.

The session length distribution shows that 53% of all sessions take less than 10 seconds, corresponding to the sessions with one filing. On the other hand, around 40% of sessions take longer than 100 seconds, which shows that rapid consumption of information is not always possible. It is worth mentioning that since we are measuring the time between the first and the last access, this does not reflect the exact session duration; the user may still be reading the last accessed filing.

We further study sessions that contain more than one filing in terms of the number of unique companies and form types that they contain. Fig. 7 shows the distributions. Most sessions are focused on a small number of companies. Specifically, 66% of all sessions contain filings from one or two companies, suggesting that EDGAR users tend to focus on filings of a single or a pair of companies in a session. The average number of unique form types in sessions is 2.78, with 25% of sessions containing a single form type, suggesting that users are interested in browsing multiple form types in a session. In case of session-based recommendation, these insights can be used to limit the search space to certain filing types from the companies accesses up until now in the current session.

We also look at the referrers to the first access in the session, which tells us how people come to interact with filings in EDGAR. It is worth mentioning that EDGAR-LFD only records accesses to filings; accesses to other pages on the SEC website are not visible in the dataset. We observe that 53% of all sessions start from an unknown referrer, including accesses from outside of the SEC website, such as web search and web links. The remaining sessions start from the various pages shown in Fig. 1a and 2: 82.5% of those start from the page containing all of the filings of a specific company (Fig. 2b), suggesting that the company page is the starting point for browsing filings for most of the sessions. Further categories of known referrers are index page (8.8%, Fig. 1a), the EDGAR archive (5.0%), filing data (3.3%), search (0.3%), and the latest filings page (0.2%). Only 0.3% of all sessions initiate from search, which shows that EDGAR users rarely use search to directly find filings; the majority of sessions start from a company page, which indicates that users probably use search to find the companies and use the company filing page (Fig. 2b) to access the filings.

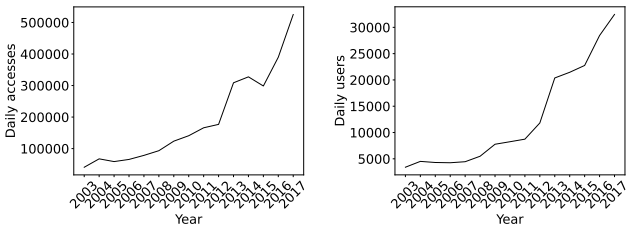


Figure 8: Average (Left:) accesses per day, and (Right:) users per day for different years.

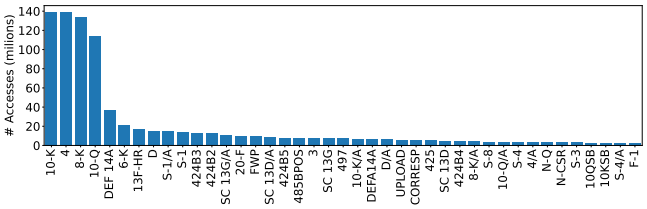


Figure 9: Most accessed filing types in the dataset.

4.3 Temporal analysis

In this part we focus on the temporal aspects of filing views. Since the data is available for more than 14 years of history, we study the changes in user access across all these years. Fig. 8 shows the average of daily accesses and daily users for each year. We observe that there is an exponential growth in both, which shows that more people are using EDGAR as a source of information year by year.

We further study the distribution of the time spent on the filings and the time between accesses and the filing dates for different filing types. We first study the number of accesses per form type. Fig. 9 shows the 41 filing types that are responsible for 90% of all accesses on EDGAR, out of 505 available form types. Over half all accesses are to four filing types, namely 4, 10-K, 10-Q, and 8-K. We look into the time spent on filings by EDGAR users. Fig. 10 (Left) shows the cumulative distribution of the time spent on filings for different filing types. Users spent the least time on filing type 4, which is essentially a table, with a median of four and 90th percentile of 100 seconds. Users spent considerably more time on 10-Q and 10-K forms; digesting information in these forms requires more time, as they report on a public company's performance over a financial year or quarter.

Fig. 10 (Right) shows the cumulative distribution of the difference between the filing date of a filing and the date that it is being viewed in days, for the top four filing types and in total. EDGAR contains filings from 1993, and EDGAR-LFD covers accesses from 2003 to 2017, so users that we study have access to filings from 10 to 24 years of history. However, 50% of accesses happen within a year of the filing dates. We further notice the difference between filing types. For filing type 4, roughly 35% of the accesses occur in a day from the filing date, and 50% happen in 12 or less days, showing that users are mostly interested in more recent filings of this type. The median is around a year for the 10-K filings, showing that 10-Ks, which cover a company's annual performance, are interesting for the users for a longer period of time.

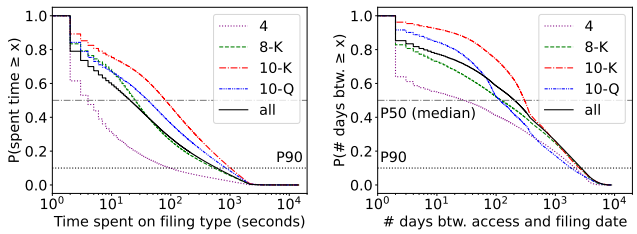


Figure 10: Cumulative distribution of (Left:) the time spent on a filing by users, for different filing types (Right) the time between access and filing date.

5 BROADER IMPLICATIONS

EDGAR is a first-source repository for analysts, investors, and regulators to access financial information [23]. EDGAR-LFD contains the interactions of users with EDGAR, in terms of the filings that they accessed. Through a user behavior analysis, we have taken the first step in understanding the information seeking behavior in the finance domain. Our goal is to provide insights that can guide the design of systems for downstream tasks.

We further highlight some of the findings in our user behavior analysis. As a use-case, we consider filing recommendation as a downstream task. With an average of over 3,000 filings being submitted per day, EDGAR users are overwhelmed with the amount of information available to them; filing recommendation reduces the burden on users to navigate the filings and facilitates quicker access overall. We would like to note that the recommendation use case presented here is just one possible task informed by our analysis. As another example, the user interactions captured in EDGAR-LFD can further be used for learning company representations that are beneficial for downstream tasks, such as identifying economically related peer firms [18].

We discuss the implications of the user behavior study for filing recommendation. The user-level analysis shows that there are two main types of users on EDGAR, less active users with a few sessions (Fig. 3 (Left)), and the top 10% more active users who account for 75% of all the sessions (Fig. 3 (Right)). More active users will benefit from different forms of filing recommendation than less active ones based on how much we know about them. For infrequent users, for whom we have no prior history beyond the current session, we consider the *next-filing recommendation* task, where the goal is to predict the next filing that a user will view, based on the filings that they have already viewed in the current session. More precisely, this is the task of recommending an accession number (see Table 2). Analogous to session-based recommendation [27], next-filing recommendation is particularly useful for EDGAR since the majority of users are not frequent users. A successful next-filing recommendation system will save users time, and will help them to have a more comprehensive picture on the subject that they are seeking information about. For more frequent users for which we have a rich history, we consider *next-session recommendation* in the context of EDGAR filings, where the session items are filing accesses during a session. The goal of the next-session recommendation task is to recommend a list of filings to the user every time they visit the website, based on the filings they have viewed in the

past. This is equivalent to sequential [30] or next basket recommendation [16]. Such recommendations would reduce the burden on users to proactively find the filings of interest every time they visit EDGAR, and facilitates quicker access to their information need. Although the frequent users are not the majority of users, they are responsible for most of the activity, justifying the design of recommendation models tailored for them. As shown in Section 4.1, the 10% most active users account for roughly 75% of the sessions on EDGAR.

The insights from our user behavior analysis can help in designing recommendation systems. Based on our session-level analysis, we know that a session usually contains filings from a small number of companies (see Fig. 7). This means that for the next-session recommendation task, we can reduce the search space tremendously by first predicting the companies that will have filings in the next session. On the other hand, the user-level analysis reveals that out of all the companies that have their filings available on EDGAR, individual users are only concerned with a small fraction of them (Fig. 4 (Left)). This means that for the next-session recommendation task, we can further limit the search space to the filings from the companies that a user has shown interest in during their past sessions. The small number of companies per session has another implication for next-filing recommendation. We can infer that a user is more likely to stick to the filings of the currently-viewed companies. In this case, the next-filing recommendation task can be reduced to ranking the filings of the companies with filings viewed in the current session. The temporal analysis demonstrates that while the filings on EDGAR go back in history as far as 24 years, EDGAR users are mostly interested in the most recent filings (Fig. 10 (Left)). This means that for ranking the filings in both recommendation scenarios, the filing date should be considered as a factor and more recent filings should have priority.

6 CONCLUSION

Financial company filings are a primary source for investors, analysts, advisors, and regulators to acquire information about a company and to support their decision making. We study the EDGAR Log File Dataset, a publicly available dataset containing the log of accesses to company filings on the US Securities and Exchange Commission (SEC) website. Through a user behavior analysis, we provide the first study on this dataset from an information access perspective. We identify two filing recommendation tasks that correspond to different usage patterns in the dataset. We find that sessions on EDGAR are focused on filings from a small number of companies and individual users are interested in a limited number of companies during their life cycle on EDGAR.

The goal of our work is to provide a stepping stone for the academic community to tackle retrieval and recommendation tasks for the finance domain. In future work, we aim to design recommendation models informed by our findings in the user behavior analysis. Moreover, the contents of the filings are available through EDGAR, and can be used to better understand the users. EDGAR-LFD contains the data for a rather long period. While we study some temporal aspects of the user behavior in this paper, many dimensions remain unexplored. For example, it will be worthwhile to see how the co-accesses to filings of different companies shift over time, and whether such a shift is a reflection of a change in companies' business lines.

ACKNOWLEDGMENTS

This research was supported by Ahold Delhaize. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

A DATA PREPARATION

Available fields. EDGAR-LFD records access to EDGAR filings between February 14, 2003 and June 30, 2017.[7] The dataset captures access to individual filings, both for the index page of a filing (Fig. 1a) and the individual documents that belong to that filing (Fig. 1b). The dataset is available in CSV format, and is based on Apache web server access logs. Next, we describe the fields captured by each EDGAR-LFD entry that are relevant to the analyses performed in this paper.[8]

- **IP** An obfuscated version of the IP address from which a filing was accessed. Obfuscation is done by replacing the last octet of the original IP address with a three character string that preserves the uniqueness of the IP address across the entire dataset.
- **Date** Date of the access in YYYY-mm-dd format.
- **Time** Time of the access in HH:MM:SS format.
- **CIK** Identifier for the company that the filing accessed in this record belongs to.
- **Accession** A unique identifier for the filing being accessed.
- **Find** A number between zero and 10 that indicates how the user arrived at the filing, e.g., internal EDGAR search.
- **Crawler** Indicates whether the user self-identifies as a web crawler.
- **Extension** The extension of the filing page being accessed. Corresponds to the "Document" column in Fig. 1a. Used to identify whether a filing document has been accessed, or its index page.

Initial cleanup. After downloading the dataset, we remove records that have an HTTP response code that indicates unsuccessful requests, which are those not in the 2xx class. We also remove records that self-identify as crawlers. We first remove entries from dates between September 23, 2005, and May 10, 2006 that were labeled by the SEC as "lost or damaged", as mentioned in [23]. We then further manually examine the days that have significantly less records than the surrounding days for no apparent reason, such as holidays. This results in marking the dates between 2003-02-14 to 2003-03-01, 2003-12-13 to 2003-12-15, and 2012-02-08 to 2012-02-29 as damaged. We remove the damaged dates from the data.

Session definition. In order to better understand the user journeys on EDGAR, we take the individual filing access records in EDGAR-LFD and group these into semantically meaningful sessions.

Following [25, 32], we define a session as a sequence of actions performed by a single IP address, where the difference in time between subsequent actions is not larger than a predefined threshold. We rely on [14] to find the threshold, based on fitting a Gaussian

[7] https://www.sec.gov/dera/data/edgar-log-file-data-set.html
[8] https://www.sec.gov/files/EDGAR_variables_FINAL.pdf for the full list of fields.

mixture model to the histogram of the time between subsequent accesses by the same IP address, i.e., same user. A three component Gaussian mixture model is fitted to the histogram using expectation maximization. After that, the point where inter-activity time is equally likely to be within the Gaussians fit with sub-hour means (within-session) and Gaussians fit with means beyond an hour (between-session) is selected as the threshold.

We apply the above method to EDGAR-LFD in order to create sessions from the individual accesses per IP address. To this end, we sample 10,000 IP addresses that we are confident come from interactions of a single human user with the system based on heuristics from [23] from each year. An IP address corresponding to a single human user is assumed to not have more than 50 requests per day. We plot the histogram of differences consecutive actions by the same IP address in seconds on a log scale. We find 35 minutes to be the optimal threshold for session breaks, which is in line with the thresholds in datasets from other domains [14].

Valid sessions. EDGAR is accessed by both humans and bots, but not all bots self-identify as such; solely relying on the `crawler` attribute of a log entry is unreliable and following previous work [23], we take additional measures to remove bot accesses. We filter out sessions based on three thresholds: (1) the number of accesses in a session, (2) the duration of a session measured in hours, and (3) the average time per access in a session. We keep the sessions that are either (a) less than four hours and have a time-per-access of more than two seconds, or (b) have less than four accesses in total.

Handling successive requests for the same filing. Each filing on EDGAR has an index page (see Fig. 1a), and files that belong to the filing. The data contains access to all documents and index pages. Our focus is on how users interact with different filings; in-filing browsing is out of scope. Hence, for each session, we ignore all successive views to the same accession number, and we remove records corresponding to access to the index pages of filings.

REFERENCES

[1] Yash Agrawal, Vivek Anand, S. Arunachalam, and Vasudeva Varma. 2021. Hierarchical Model for Goal Guided Summarization of Annual Financial Reports. In *Companion of The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*. ACM / IW3C2, 247–254.

[2] Qingyao Ai, Susan T. Dumais, Nick Craswell, and Dan Liebling. 2017. Characterizing Email Search Using Large-Scale Behavioral Logs and Surveys. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1511–1520.

[3] Baptiste Barreau and Laurent Carlier. 2020. History-Augmented Collaborative Filtering for Financial Recommendations. In *RecSys 2020: Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, September 22-26, 2020*. ACM, 492–497.

[4] Horatiu Bota, Adam Fourney, Susan T. Dumais, Tomasz L. Religa, and Robert Rounthwaite. 2018. Characterizing Search Behavior in Productivity Software. In *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, CHIIR 2018, New Brunswick, NJ, USA, March 11-15, 2018*. ACM, 160–169.

[5] Diego Ceccarelli, Francesco Nidito, and Miles Osborne. 2016. Ranking Financial Tweets. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*. ACM, 527–528.

[6] Jun Chang, Wenting Tu, Changrui Yu, and Chuan Qin. 2021. Assessing dynamic qualities of investor sentiments for stock recommendation. *Inf. Process. Manag.* 58, 2 (2021), 102452.

[7] Chi Chen, Li Zhao, Jiang Bian, Chunxiao Xing, and Tie-Yan Liu. 2019. Investment Behaviors Can Tell What Inside: Exploring Stock Intrinsic Properties for Stock Trend Prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*. ACM, 2376–2384.

[8] Dawei Cheng, Fangzhou Yang, Xiaoyang Wang, Ying Zhang, and Liqing Zhang. 2020. Knowledge Graph-based Event Embedding Framework for Financial Quantitative Investments. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*. ACM, 2221–2230.

[9] Michael S. Drake, Bret A. Johnson, Darren T. Roulstone, and Jacob R. Thornock. 2020. Is There Information Content in Information Acquisition? *The Accounting Review* 95, 2 (2020), 113–139.

[10] Michael S. Drake, Darren T. Roulstone, and Jacob R. Thornock. 2015. The Determinants and Consequences of Information Acquisition via EDGAR. *Contemporary Accounting Research* 32, 3 (2015), 1128–1161.

[11] Mengzhen Fan, Dawei Cheng, Fangzhou Yang, Siqiang Luo, Yifeng Luo, Weining Qian, and Aoying Zhou. 2020. Fusing Global Domain Information and Local Semantic Information to Classify Financial Documents. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*. ACM, 2413–2420.

[12] Fuli Feng, Moxin Li, Cheng Luo, Ritchie Ng, and Tat-Seng Chua. 2021. Hybrid Learning to Rank for Financial Event Ranking. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. ACM, 233–243.

[13] Fuli Feng, Cheng Luo, Xiangnan He, Yiqun Liu, and Tat-Seng Chua. 2020. FinIR 2020: The First Workshop on Information Retrieval in Finance. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*. ACM, 2451–2454.

[14] Aaron Halfaker, Oliver Keyes, Daniel Kluver, Jacob Thebault-Spieker, Tien T. Nguyen, Kenneth Shores, Anuradha Uduwage, and Morten Warncke-Wang. 2015. User Session Identification Based on Strong Regularities in Inter-activity Time. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*. ACM, 410–418.

[15] James A. Hodson and James Y. Zhang. 2014. Entity extraction and disambiguation in finance. In *ERD'14, Proceedings of the First ACM International Workshop on Entity Recognition & Disambiguation, July 11, 2014, Gold Coast, Queensland, Australia*. ACM, 1–2.

[16] Haoji Hu and Xiangnan He. 2019. Sets2Sets: Learning from Sequential Sets with Neural Networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*. ACM, 1491–1499.

[17] Laura Korkeamäki and Sanna Kumpulainen. 2019. Interacting with Digital Documents: A Real Life Study of Historians' Task Processes, Actions and Goals. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR 2019, Glasgow, Scotland, UK, March 10-14, 2019*. ACM, 35–43.

[18] Charles M.C. Lee, Paul Ma, and Charles C.Y. Wang. 2015. Search-based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics* 116, 2 (2015), 410 – 431.

[19] F. Li and Chengzhu Sun. 2018. Information acquisition and expected returns: Evidence from EDGAR search traffic. *Social Science Research Network* (2018), 1.

[20] Ruihai Li, Xuewu (Wesley) Wang, Zhipeng Yan, and Yan Zhao. 2019. Sophisticated Investor Attention and Market Reaction to Earnings Announcements: Evidence From the SEC's EDGAR Log Files. *Journal of Behavioral Finance* 20, 4 (2019), 490–503.

[21] Jiqun Liu, Matthew Mitsui, Nicholas J. Belkin, and Chirag Shah. 2019. Task, Information Seeking Intentions, and User Behavior: Toward A Multi-level Understanding of Web Search. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR 2019, Glasgow, Scotland, UK, March 10-14, 2019*. ACM, 123–132.

[22] Yu-Wen Liu, Liang-Chih Liu, Chuan-Ju Wang, and Ming-Feng Tsai. 2016. FIN10K: A Web-based Information System for Financial Report Analysis and Visualization. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*. ACM, 2441–2444.

[23] Tim Loughran and Bill McDonald. 2017. The Use of EDGAR Filings by Investors. *Journal of Behavioral Finance* 18, 2 (2017), 231–248.

[24] Zhiqiang Ma, Steven Pomerville, Mingyang Di, and Armineh Nourbakhsh. 2020. SPot: A Tool for Identifying Operating Segments in Financial Tables. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*. ACM, 2157–2160.

[25] Gilad Mishne and Maarten de Rijke. 2006. A Study of Blog Search. In *Advances in Information Retrieval*. Springer Berlin Heidelberg, Berlin, Heidelberg, 289–301.

[26] Vassilis Plachouras, Charese Smiley, Hiroko Bretz, Ola Taylor, Jochen L. Leidner, Dezhao Song, and Frank Schilder. 2016. Interacting with Financial Data using Natural Language. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*. ACM, 1121–1124.

[27] Pengjie Ren, Zhumin Chen, Jing Li, Zhaochun Ren, Jun Ma, and Maarten de Rijke. 2019. RepeatNet: A Repeat Aware Neural Recommendation Machine for Session-Based Recommendation. In *The Thirty-Third AAAI Conference on*

Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. AAAI Press, 4806–4813.

[28] Ming-Feng Tsai and Chuan-Ju Wang. 2013. Risk Ranking from Financial Reports. In *Advances in Information Retrieval - 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings (Lecture Notes in Computer Science, Vol. 7814)*. Springer, 804–807.

[29] Pingle Wang. 2019. Demand for Information and Stock Returns: Evidence from EDGAR. *Available at SSRN 3348513* (2019).

[30] Shoujin Wang, Liang Hu, Yan Wang, Longbing Cao, Quan Z. Sheng, and Mehmet A. Orgun. 2019. Sequential Recommender Systems: Challenges, Progress and Prospects. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. ijcai.org, 6332–6338.

[31] Xiaochuan Wang, Ning Su, Zexue He, Yiqun Liu, and Shaoping Ma. 2018. A Large-Scale Study of Mobile Search Examination Behavior. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 1129–1132.

[32] Wouter Weerkamp, Richard Berendsen, Bogomil Kovachev, Edgar Meij, Krisztian Balog, and Maarten de Rijke. 2011. People searching for people: analysis of a people search engine log. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*. ACM, 45–54.

[33] Yiying Yang, Zhongyu Wei, Qin Chen, and Libo Wu. 2019. Using External Knowledge for Financial Event Prediction Based on Graph Neural Networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*. ACM, 2161–2164.

[34] Xiaoting Ying, Cong Xu, Jianliang Gao, Jianxin Wang, and Zhao Li. 2020. Time-aware Graph Relational Attention Network for Stock Recommendation. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*. ACM, 2281–2284.