Received: 3 June 2021 | Revised: 11 February 2022 | Accepted: 6 March 2022 | Published on: 15 March 2022

DOI: 10.1002/smj.3397

RESEARCH ARTICLE

STRATEGIC MANAGEMENT JOURNAL WILEY

Revealing the revealed preferences of public firm CEOs and top executives: A new database from credit card spending

Joseph Raffiee | **Daniel Fehder** | **Florenta Teodoridis**

Department of Management and Organization, Marshall School of Business, University of Southern California, Los Angeles, California, USA

Correspondence
Joseph Raffiee, Department of Management and Organization, Marshall School of Business, University of Southern California, 701 Exposition Blvd., HOH 512, Los Angeles, CA 90089, USA.
Email: joe.raffiee@marshall.usc.edu

Abstract

Research Summary: We introduce a new database which provides an unprecedented window into the off-the-job lives and interests of public firm top executives as reflected by their personal income allocation. We construct this database by matching household credit card spending data with the population of executives in Execucomp. To overcome the significant computational challenges associated with matching these data, we build on the statistical record-linking literature in a way that allows us to generate reliable matches with limited information. To facilitate research exploring new questions made possible with this database, we make our matching crosswalk freely available for academic use.

Managerial Summary: This article describes the matching procedures associated with the development of a database which sheds new light on the revealed preferences of public firm top executives. Prior work on upper echelons has routinely stressed how preferences and characteristics of top executives often manifest in firm behaviors and are important predictors of firm outcomes. Nevertheless, the lack of a reliable paper

All authors contributed equally and are listed in a randomized order.

trail capturing executive preferences, particularly at a large scale, has been a friction slowing comprehensive empirical research on this topic. We describe how we address this limitation by linking top executives listed in the database Execucomp with credit card spending data provided by the consumer data provider L2 and outline a number of new research questions made possible by our matching efforts. The results of our matching efforts (Execucomp-L2 unique identifier crosswalk) are available at https://github.com/dfehder/FRT_2022_SMJ_CEO_TopExec_Preferences.

KEYWORDS

probabilistic matching, record linkage, revealed preferences, top management teams, upper echelons theory

1 | INTRODUCTION

In 1984, Hambrick and Mason introduced upper echelons theory (UET) based on the premise that organizations are reflections of their top executives and hence "If we want to understand why organizations do the things they do, or why they perform the way they do, we must consider the biases and dispositions of their most powerful actors – their top executives" (Hambrick, 2007, p. 334). With UET as the conceptual backdrop, the decades which followed gave rise to a substantial body of research demonstrating that strategic actions, decisions, and ultimately firm outcomes can be predicted by the characteristics of a firm's top executives (for reviews, see Cannella, Finkelstein, & Hambrick, 2008; Carpenter, Geletkanycz, & Sanders, 2004; Neely, Lovelace, Cowen, & Hiller, 2020; Wang, Holmes, Oh, & Zhu, 2016).

The growth in upper echelons empirical research has been enabled in part by accessible data. Commercial databases such as Execucomp are readily available and allow scholars to measure a host of observable managerial characteristics originally theorized by Hambrick and Mason (1984). These data are easily linked to firm-level databases such as Compustat via common unique identifiers.[1] To move beyond observables, upper echelons scholars have leveraged a number of other data sources in creative ways. For example, researchers have used public political contribution records to infer executive political ideology (e.g., Chin, Hambrick, & Treviño, 2013; Chin & Semadeni, 2017; Christensen, Dhaliwal, Boivie, & Graffin, 2015; Graffin, Hubbard, Christensen, & Lee, 2020) and employed methodological advancements in text, historiometric, and videographic analysis to analyze public documents and media content (e.g., earnings call transcripts, shareholder letters, news articles, video clips, etc.) in efforts to penetrate the managerial "black box" (Lawrence, 1997) and create proxies for executive

[1]Execucomp and Compustat are widely used commercial data products, with Execucomp providing detailed information of public firm top executives and Compustat providing detailed information on public firms. There is a common crosswalk between these data sources, allowing scholars to easily link executives to firms.

personality and/or cognition (e.g., Choudhury, Wang, Carlson, & Khanna, 2019; Gamache, Neville, Bundy, & Short, 2020; Harrison, Thurgood, Boivie, & Pfarrer, 2019; Hill, Recendes, & Ridge, 2019; Wowak, Mannor, Arrfelt, & McNamara, 2016).

While these studies have significantly advanced our understanding of UET (see Neely et al., 2020, for a recent review), an emergent body of research has begun to explore how characteristics of executives' off-the-job lives—their interests, behaviors, preferences, and hobbies—may manifest and predict a variety of organizational outcomes/stakeholder evaluations (Cronqvist, Makhija, & Yonker, 2012; Davidson, Dey, & Smith, 2015; Sunder, Sunder, & Zhang, 2017). As recently described by Ouyang et al. (2021, pp. 1–2), "Examining executives' personal lives outside of the workplace is important because like everyone else, executives' cognitive base and value systems are shaped by their activities both on and off the job." The challenge, of course, is that systematically and reliably observing characteristics of executives' off-the-job lives, particularly at a large scale, is a formidable task, as executives do not routinely report their off-the-job interests and activities to a common source. As a result, emerging work in this area has tended to focus on a limited set of off-the-job activities and interests, such as flying a personal aircraft (Cain & McKeon, 2016; Ouyang et al., 2021; Sunder et al., 2017), purchasing real estate (Davidson et al., 2015), golfing (Biggerstaff, Cicero, & Puckett, 2017), or participation in marathons (Campbell & Zipay, 2019; Limbach & Sonnenburg, 2015)—with the collection of such data typically involving resource-intensive and painstakingly time-consuming (and often manual) searches of public records.

In this article, we hope to further catalyze upper echelons empirical research and facilitate research exploring executive nonworkplace activity by introducing and constructing a new dataset which provides an unprecedented and detailed window into the off-the-job lives, preferences, and interests of U.S. public firm executives. The data we introduce are sourced directly from the credit reporting agency Experian and are derived from household credit card spending patterns. Hence, these data capture revealed preferences and interests based on income allocation (Samuelson, 1938). The data we use are provided by the commercial data provider L2, who sources data from Experian and other high-quality data providers and includes 94 fields of consumer and household information pertaining to over 180 million Americans.[2] These data capture a wide array of personal interests ranging from sports and collectibles to fine dining and charitable giving. L2 has been a highly trusted and leading source of voter information and consumer data for more than 50 years, providing data to congressional and state offices, for-profit firms, consultants, media outlets, and universities, among others. To facilitate upper echelons research that leverages these data, we match the L2 data to the population of executives listed in Execucomp, thereby providing a novel and hitherto unparalleled view into the nonworkplace interests and lives of top executives. The matched data can be easily linked to additional firm-level databases such as Compustat using existing Compustat-Execucomp unique identifiers. We make the

[2]In addition, these data also include voter registration records and voter history records for state, local, and federal elections, which are sourced directly from state records. While we emphasize the novelty of the credit card-derived revealed preferences, voter registration and voter history records provide a complement to existing donation-based proxies commonly employed to infer executive ideology in extant strategy research (e.g., Chin et al., 2013). Our matched crosswalk will allow future researchers to use these data in addition to the revealed preferences credit card data. The matched crosswalk and the underlying code producing it are available at the following GitHub repository: https://github.com/dfehder/FRT_2022_SMJ_CEO_TopExec_Preferences.

results from our matching procedure (i.e., L2-Execucomp crosswalk) free and publicly available for academic use.[3]

The central and substantial friction we face when constructing our dataset is the problem of record-linking with limited information, a common issue when linking disparate sources of archival data. Despite providing detailed information on individual preferences and interests, the L2 data contain limited information about individuals other than names, demographics, and home addresses. This creates a serious challenge when attempting to link L2 records with records of executives listed in Execucomp using prevailing record-matching techniques, which typically rely on fuzzy string matching using first and last names along with at least one additional field common across both datasets. For example, existing work linking executives to political contribution records typically utilizes fuzzy (or exact) string matching on a combination of first and last names *and* the employer (e.g., Graffin et al., 2020). This additional field or discriminating variable plays a central role in the matching process as it theoretically limits the consideration set of potential matches to a small enough number of likely matches where potential false positives are assumed to be rare enough that any links are considered credible. We lack such a variable across L2 and Execucomp.

To overcome this challenge, our approach builds on the computer science and statistical record-linking literature in a way that allows us to generate reliable matches with limited information and without a high-quality discriminating variable such as employer (Christen, 2012). Specifically, our statistical name-matching strategy is predicated on our ability to uncover and utilize implicit heterogeneity in names and location by integrating external data sources that allow us to infer how rare or infrequent specific combinations of names and locations are in the overall population. Our matching procedure works by identifying (a) matches and (b) weighting these matches by the probability of observing such a match at random (Christen, 2012; Fellegi & Sunter, 1969; Winkler, 1990). In addition, we address computational challenges driven by the very large number of records in L2 (approximately 180 million records) by reducing the dimensionality of the data and parallel processing the matching procedure. Absent these approaches, we estimate common fuzzy (or exact) string matching techniques to require roughly 2 million hours of run time to evaluate the more than 7 trillion comparisons needed to match records across L2 and Execucomp.

Using this procedure, we are able to match 26,257 executives from Execucomp to L2 records. Our matching procedure produces a continuous posterior probability score for each match which reflects the probability that the match is correct. More than 80% of the matches identified by our approach received a posterior probability score equal to or greater than 95%. We further verify the reliability of our record-linking strategy by benchmarking our results against results from prevailing record-linking procedures when names *and* employer are available across datasets. This exercise validates our approach by demonstrating that our matching results compare favorably with prevailing fuzzy string matching procedures that match on individual names and rely on employer as a discriminating variable. Our resulting matched crosswalk between datasets is made publicly available for academic use.

[3]Using our crosswalk requires researchers (or their institution) to obtain a license from L2 (https://l2-data.com/our-data/) which provides access to the underlying data, in the same way that Execucomp and Compustat require data licenses for their use. We obtained the license via L2's contact page and asked for their academic data pricing package.

This article makes two primary contributions to the strategy and upper echelons literature. First, we provide a new dataset which provides an unprecedented window into the nonwork lives and interests of public firm top executives in the United States, as revealed by household credit card expenditures. By making the resulting data from our matching procedure publicly available, we contribute to the creation of a public good which we hope will reduce frictions for researchers wishing to study how the characteristics of executive off-the-job lives may relate to a broad array of organizational aspects. Second, we provide a discussion of new research questions and research directions that can be explored using our data to advance UET, including but not limited to deductive theory-driven testing, machine learning approaches for pattern recognition, the construction of composite measures based on deductive theory, and the integration of new data sources to further leverage the insights our revealed preferences variables and statistical matching procedures make possible.

The remainder of this article is organized as follows. We begin by introducing the L2 data and describing its contents. We then describe and implement our record-linking strategy which allows us to match the L2 data with Execucomp despite limited overlapping information. We conclude with a discussion of our contributions and future research.

2 | A NOVEL DATASET OF OFF-THE-JOB EXECUTIVE CHARACTERISTICS

2.1 | The L2 consumer dataset

The primary data source we use is obtained from L2, a leading commercial consumer data supplier for political campaigns, trade groups and associations, academic organizations, analytics firms, and media/digital consultants. L2 provides rich information about individuals derived from public and commercial data sources, with consumer data sourced directly from credit card spending data provided by Experian, a leading credit reporting agency. L2 invests significant resources to ensure data accuracy and reliability, including rigorous data hygiene protocols, data verification, and data cross-indexing. For these reasons, L2 has been a highly trusted and leading data source for more than 50 years, providing voter information and consumer data to more congressional and state offices than any other data supplier, while also supplying data to for-profit firms, consultants, and other buyers. The data provided by L2 have been licensed by numerous universities and research institutions such as The Pew Charitable Trusts, along with dozens of Media outlets including the *New York Times*, CBS, ABC, NPR, Fox, the *Washington Post*, and Univision, among others.

The L2 data contain rich information on consumer and household characteristics, voter registration, voting history, and current demographic information for more than 180 million adults in the United States. In this article, we focus primarily on the consumer and household data provided by L2 which is primarily sourced directly from the credit reporting agency Experian, as these data provide a novel window into the behavioral interests of individuals as derived from household credit card spending.[4] A complete list of consumer and household data fields

[4]There are three major credit bureaus: Experian, Equifax, and TransUnion. While not all financial transactions are reported to all three bureaus (i.e., a creditor may not report a delinquency to all bureaus), we have no reason to believe that such reporting is systematically biased. While credit scores across bureaus are unlikely to be identical, they correlate highly.

TABLE 1 Descriptions and summary statistics of L2 variables matched to executives with a 95% or better posterior match probability

Variable	Observation level	L2's source	Mean	*SD*	Min	Max
PetOwner_Cat	Household	Experian	0.05	0.22	0.0	1.0
PetOwner_Dog	Household	Experian	0.10	0.30	0.0	1.0
PetOwner_Horse	Household	Experian	0.02	0.15	0.0	1.0
PetOwner_Other	Household	Experian	0.19	0.39	0.0	1.0
BookBuyer	Household	Experian	0.77	1.46	0.0	9.0
Interest_in_Current_Affairs_Politics	Household	Experian	0.22	0.41	0.0	1.0
Buyer_Antiques	Household	Experian	0.00	0.04	0.0	1.0
Buyer_Art	Household	Experian	0.26	0.44	0.0	1.0
Interest_in_Automotive_Parts_Accessories	Household	Experian	0.14	0.35	0.0	1.0
CulinaryInterestMagazine	Household	Experian	0.05	0.25	0.0	4.0
DoItYourselferMagazine	Household	Experian	0.09	0.42	0.0	7.0
DonatesEnvironmentCause	Household	Experian	0.03	0.18	0.0	1.0
DonatesToCharity	Household	Experian	0.54	0.50	0.0	1.0
FamilyMagazine	Household	Experian	0.32	0.77	0.0	9.0
FemaleOrientedMagazine	Household	Experian	0.06	0.28	0.0	4.0
FinancialMagazine	Household	Experian	0.24	0.82	0.0	8.0
GardeningMagazine	Household	Experian	0.06	0.36	0.0	5.0
HealthFitnessMagazine	Household	Experian	0.69	1.38	0.0	9.0
PoliticalContributer	Household	Experian	0.55	0.75	0.0	9.0
ReligiousMagazine	Household	Experian	0.00	0.02	0.0	1.0
Collector_Antiques	Household	Experian	0.29	0.45	0.0	1.0
Collector_Arts	Household	Experian	0.03	0.17	0.0	1.0
Collector_Avid	Household	Experian	0.03	0.18	0.0	1.0
Collector_Coins	Household	Experian	0.03	0.17	0.0	1.0
Collector_General	Household	Experian	0.32	0.47	0.0	1.0
Collector_Military	Household	Experian	0.01	0.07	0.0	1.0
Collector_Sports	Household	Experian	0.04	0.19	0.0	1.0
Collector_Stamps	Household	Experian	0.01	0.12	0.0	1.0
Donates_to_Animal_Welfare	Individual	Experian	0.03	0.16	0.0	1.0
Donates_to_Arts_and_Culture	Individual	Experian	0.00	0.07	0.0	1.0
Donates_to_Childrens_Causes	Individual	Experian	0.04	0.19	0.0	1.0
Donates_to_Conservative_Causes	Individual	Experian	0.00	0.04	0.0	1.0
Donates_to_Healthcare	Individual	Experian	0.15	0.35	0.0	1.0
Donates_to_International_Aid_Causes	Individual	Experian	0.00	0.03	0.0	1.0
Donates_to_Liberal_Causes	Individual	Experian	0.00	0.03	0.0	1.0
Donates_to_Local_Community	Individual	Experian	0.23	0.42	0.0	1.0

TABLE 1 (Continued)

Variable	Observation level	L2's source	Mean	*SD*	Min	Max
Donates_to_Wildlife_Preservation	Individual	Experian	0.06	0.23	0.0	1.0
Donates_to_Veterans_Causes	Individual	Experian	0.02	0.14	0.0	1.0
Interest_in_Auto_Work	Household	Experian	0.30	0.46	0.0	1.0
Interest_in_Aviation	Household	Experian	0.02	0.13	0.0	1.0
Interest_in_BoardGames_Puzzles	Household	Experian	0.00	0.07	0.0	1.0
Interest_in_Boating_Sailing	Household	Experian	0.22	0.41	0.0	1.0
Interest_in_Camping_Hiking	Household	Experian	0.11	0.31	0.0	1.0
Interest_in_Cooking_General	Household	Experian	0.52	0.50	0.0	1.0
Interest_in_Cooking_Gourmet	Household	Experian	0.53	0.50	0.0	1.0
Interest_in_Crafts	Household	Experian	0.47	0.50	0.0	1.0
Interest_in_Education_Online	Household	Experian	0.05	0.22	0.0	1.0
Interest_in_Electronic_Gaming	Household	Experian	0.03	0.17	0.0	1.0
Interest_in_Exercise_Aerobic	Household	Experian	0.04	0.19	0.0	1.0
Interest_in_Exercise_Health	Household	Experian	0.55	0.50	0.0	1.0
Interest_in_Exercise_Running_Jogging	Household	Experian	0.02	0.15	0.0	1.0
Interest_in_Exercise_Walking	Household	Experian	0.07	0.25	0.0	1.0
Interest_in_Fishing	Household	Experian	0.10	0.29	0.0	1.0
Interest_in_Food_Wines	Household	Experian	0.05	0.21	0.0	1.0
Interest_in_Foods_Natural	Household	Experian	0.11	0.31	0.0	1.0
Interest_in_Gaming_Casino	Household	Experian	0.02	0.14	0.0	1.0
Interest_in_Gardening	Household	Experian	0.65	0.48	0.0	1.0
Interest_in_Golf	Household	Experian	0.19	0.39	0.0	1.0
Interest_in_History_Military	Household	Experian	0.04	0.19	0.0	1.0
Interest_in_Home_Furnishings	Household	Experian	0.74	0.44	0.0	1.0
Interest_in_Home_Improvement	Household	Experian	0.24	0.43	0.0	1.0
Interest_in_Home_Repair	Household	Experian	0.02	0.15	0.0	1.0
Interest_in_House_Plants	Household	Experian	0.00	0.00	0.0	0.0
Interest_in_Hunting	Household	Experian	0.02	0.15	0.0	1.0
Interest_in_Motorcycling	Household	Experian	0.03	0.17	0.0	1.0
Interest_in_Musical_Instruments	Household	Experian	0.04	0.19	0.0	1.0
Interest_in_Nascar	Household	Experian	0.05	0.22	0.0	1.0
Interest_in_Photography	Household	Experian	0.08	0.27	0.0	1.0
Interest_in_Photography_Video	Household	Experian	0.01	0.10	0.0	1.0
Interest_in_Religious_Inspirational	Household	Experian	0.15	0.35	0.0	1.0
Interest_in_Science_Space	Household	Experian	0.04	0.18	0.0	1.0
Interest_in_Scuba_Diving	Household	Experian	0.01	0.08	0.0	1.0

TABLE 1 (Continued)

Variable	Observation level	L2's source	Mean	*SD*	Min	Max
Interest_in_Sewing_Knitting	Household	Experian	0.07	0.25	0.0	1.0
Interest_in_Shooting	Household	Experian	0.27	0.44	0.0	1.0
Interest_in_Smoking	Household	Experian	0.03	0.16	0.0	1.0
Interest_in_Snow_Skiing	Household	Experian	0.02	0.15	0.0	1.0
Interest_in_SpectatorSports_Auto_Racing	Household	Experian	0.03	0.16	0.0	1.0
Interest_in_SpectatorSports_Baseball	Household	Experian	0.08	0.27	0.0	1.0
Interest_in_SpectatorSports_Basketball	Household	Experian	0.07	0.25	0.0	1.0
Interest_in_SpectatorSports_Football	Household	Experian	0.09	0.29	0.0	1.0
Interest_in_SpectatorSports_Hockey	Household	Experian	0.03	0.18	0.0	1.0
Interest_in_SpectatorSports_on_TV	Household	Experian	0.06	0.24	0.0	1.0
Interest_in_SpectatorSports_Soccer	Household	Experian	0.01	0.07	0.0	1.0
Interest_in_Sports_Leisure	Household	Experian	0.40	0.49	0.0	1.0
Interest_in_Sweepstakes_Contests	Household	Experian	0.17	0.37	0.0	1.0
Interest_in_Tennis	Household	Experian	0.01	0.10	0.0	1.0
Interest_in_the_Arts	Household	Experian	0.31	0.46	0.0	1.0
Interest_in_Theater_Performing_Arts	Household	Experian	0.11	0.32	0.0	1.0
Interest_in_Travel_Cruise	Household	Experian	0.09	0.28	0.0	1.0
Interest_in_Travel_Domestic	Household	Experian	0.38	0.49	0.0	1.0
Interest_in_Travel_International	Household	Experian	0.06	0.23	0.0	1.0
Interest_in_Woodworking	Household	Experian	0.04	0.19	0.0	1.0
Gun_Owner	Individual	Experian	0.22	0.42	0.0	1.0
Veteran	Individual	State records	0.06	0.23	0.0	1.0

and corresponding summary statistics for the population of executives we match, as described below, are provided in Table 1. A full list of available fields in the data, including voter registration and voter history, are provided in Appendix A.[5]

As detailed in Table 1, L2 supplies researchers with a number of fields which indicate interests as reflected by credit card purchases within the household. As an example, an individual would be flagged as having an interest in international travel if their household credit card expenditures reflect a history of international flight purchases and/or credit card spending in international locations. Likewise, an individual would be flagged as having an interest in auto

[5]Voter registration and voter history data are sourced directly from state records. There is variation, however, with respect to the information contained in voter registration records, specifically whether or not state records contain partisan affiliation. For states where voter registration records do not contain partisan affiliation (e.g., open primary states), L2 provides a predicted partisan affiliation using a proprietary partisan classification algorithm which incorporates voter turnout history and the consumer data in their prediction models. Our matched data crosswalk will allow researchers to use all data provided by L2 (e.g., voter registration), not just the consumer and household data, which is the primary focus of this article.

work if household credit card spending reflects a history of auto work-related purchases.[6] As reflected in Table 1, the level of granularity varies by field. In some cases, the field reflects a broad interest, such as an interest in exercise health. In other fields, the data reflect specific and targeted behaviors, such as a subscription to a health and fitness magazine.

The purpose of these data is to facilitate political and marketing campaigns by empowering political operatives, marketers, and consultants to target consumers based on their interests. As a result, when the data indicate that an individual has a given interest (e.g., an interest in Nascar), there is certainty that the interest exists based on spending patterns. In other words, an affirmative indicator for the consumer interest data reflects certainty from L2 that the interest is present. In contrast, a lack of an indication that there is an interest in a given consumer field does not mean that L2 is certain that such an interest does not exist, but rather that they do not have enough spending data to conclusively conclude that the interest is present. Hence, a lack of an affirmative indicator reflects that L2 cannot make a conclusive affirmative conclusion.

2.2 | Linking L2 with Execucomp

2.2.1 | Matching challenges

While the primary use of these data heretofore has been to aid the development of targeted marketing campaigns, when matched to Execucomp, the L2 data provide researchers with an unprecedented and large-scale view into the off-the-job lives and interests of U.S. public firm top executives and hence a treasure trove of new and otherwise difficult to observe measures of executive interests. That said, matching these data to Execucomp (or other data sources) represents a formidable challenge because there is limited overlap of fields between the L2 data and Execucomp other than the first and the last name. Most notably, the L2 data does not include information on employer, a piece of information commonly used in conjunction with individual names to link records across databases in social-science research (Chin & Semadeni, 2017). The intuition behind this prevailing approach is that matching on name *and* employer reduces the probability of false positives as would likely be the case if matching on name alone. Hence, to utilize the L2 data, we need a way to credibly match the names of executives listed in Execucomp with those listed in the L2 data.

2.2.2 | Our matching strategy

While the focus of our approach is to reliably match the L2 data with Execucomp to facilitate substantive research which leverages these data, we note that the method we present can be applied to other data matching tasks in management and social-science research. Scholars often

[6]Many of the fields provided by L2 are at the household level (noted in Table 1 by the Observation level column). Within a family, Experian cannot resolve who made the actual purchase if multiple cards are linked in the account (e.g., typically spouses with access to a joint credit card account). Thus, an interest in auto work is observed by spending on a credit card directly linked to the individual, but we cannot resolve with certainty if the interest in auto work might be attributed to the CEO or the CEO's spouse. While we keep this in mind when interpreting the data, it is noteworthy that empirical evidence suggests there tends to be strong spousal similarity across a wide variety of dimensions (Buss, 1985).

need to link different data sources when conducting empirical research. To do so, the prevailing methods typically require a number of overlapping fields across both data sources. For example, and as described above, when matching political contribution data to top executives in Execucomp (e.g., Chin & Semadeni, 2017), researchers typically implement fuzzy string matching using names and at least one other field present in both data sources (a discriminating variable), often times the employer. This additional field plays a central role in the matching process, as it theoretically limits the consideration set of potential matches to a small enough number of likely matches where potential false positives are assumed to be rare enough that any links are considered credible. In other words, when matching records from two data sources, the probability of a false positive (incorrect match) falls when two records share the same name *and* same employer, relative to two records which share the same name only.

To credibly match the L2 data with Execucomp, we must overcome the challenge arising from the fact that a clear discriminating variable, such as employer, is not available in both data sets. To do so, we turn to the record linkage literature which has roots in the fields of computer science and mathematical statistics (Christen, 2012). Specifically, our approach leverages the theoretical insights of probabilistic matching pioneered by Fellegi and Sunter (1969) and further developed by Enamorado, Fifield, and Imai (2019). The core intuition of this approach is that, even in the absence of a high-quality discriminating variable, a number of individual lower quality overlapping variables, when considered jointly, will allow us to identify matches across datasets with high certainty. A number of other fuzzy matching algorithms also implement a method akin probabilistic record-matching technique (Wasi & Flaaen, 2015). However, these approaches rely heavily on user's private knowledge about characteristics of the lower quality overlapping variables, such as frequency of combinations. The Enamorado et al. (2019) algorithm we employ in this article relies on more objective cues derived from data distribution attributes.[7] Figure 1 documents the steps we take in a flowchart that can be followed to replicate our data or to apply the same technique to other datasets.

More specifically, to implement this approach, we focus on overlapping variables of first name, last name, gender, age, income, and commuting region. We make use of the U.S. Census's definitions of Core-Based Statistical Areas as counties which have economic and social relationships defined mainly by the commuting behaviors across counties either to or from work. Because these definitions use employee commuting behavior to define regions, we believe that they form a good measure of the "risk set" of places a company's top management team (TMT) are likely to live.[8] These variables are provided in both L2 and Execucomp. As opposed to straight one-to-one matching on these variables, the approach of Enamorado et al. (2019) returns a list of potential matches, a probability estimate for each match in terms of correctness, and a set of statistics that provide information about the success quality of the overall matching process. The success and probability estimates for the overall matching process and for each individual match are determined by an algorithmic process that takes into account the frequency and commonness of variables used in the matching process.

[7]A detailed explanation of the precise mechanics of this technique and comparisons with other algorithms are outside the scope of this article. We refer readers to Christen (2012) and Herzog, Scheuren, and Winkler (2007) which provide excellent primers.

[8]Details about our regional definitions and the process of how we include these in our matching algorithm are included in Appendix B.

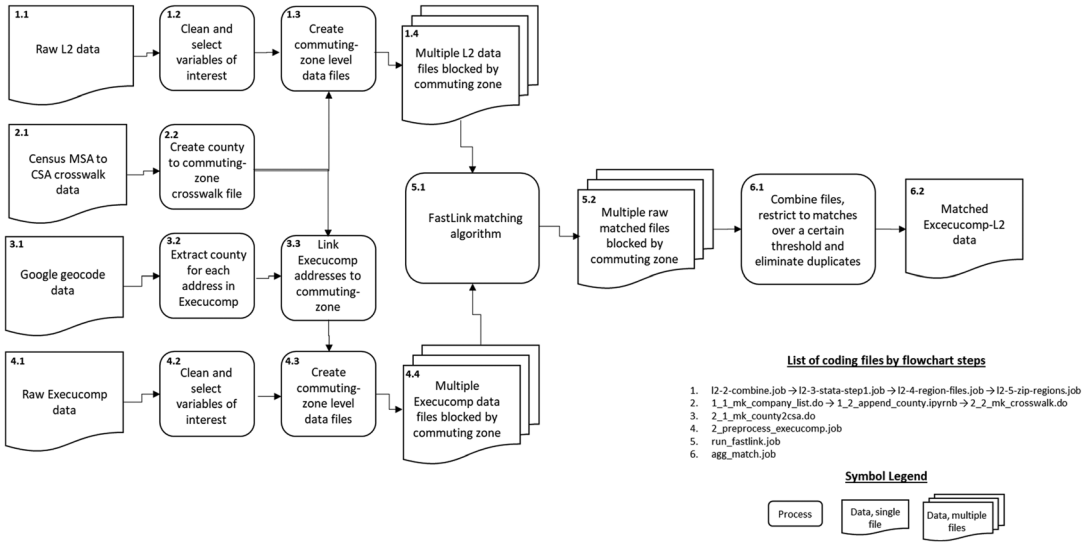

FIGURE 1 L2 to Execucomp matching flowchart

To demonstrate the intuition, a potential match that shares the last name Smith would be given less weight than a potential match that shares the last name Hambrick, as Hambrick is a less common last name compared to Smith in the U.S. population. The Enamorado et al. (2019) algorithm implements this logic for each variable used in the matching process and simultaneously determines the weights assigned to each variable (e.g., names receive more weight than gender) using the expectation maximization algorithm while also placing more weight on matches which are in infrequent parts of the distribution (e.g., infrequent names receive more weight than frequent names). The output of this process is a continuous composite probability score for each potential match that allows researchers to specify the level of certainty required to consider a match correct (i.e., the acceptable trade-off between false positives and false negatives given the researcher's topic of interest).

We apply this method to match the L2 data with Execucomp. As described above, the L2 data consist of records for over 180 million (precisely, 182,464,067) adults in the United States. Our sample of executives consists of executives listed in Execucomp employed by public firms headquartered in the United States between the years of 2000 and 2019. This reflects 40,710 unique executives working for 3,128 companies. The size of the two data sets presents a number of computational challenges. To illustrate, the number of comparisons which need to be made to implement the Enamorado et al. (2019) algorithm directly is over 7 trillion, as the number of comparisons necessary when matching two data sources is the product of the size of the two original data sets. Assuming each comparison takes a millisecond—a typical comparison time when using common fuzzy string matching techniques—it would take roughly 2 million hours of run time (or approximately 83,000 days or 228 years) to compare all records across our two datasets. Even if we were to engage only in exact matching, which would allow comparison times to approach speeds as fast as a microsecond per comparison, it would still take roughly 83 days of run time. We address these computational challenges in two ways: reducing the dimensionality of the data and parallel processing of the matching procedure.

To reduce the dimensionality of the data, we employ the most commonly accepted practice in the record linkage literature: blocking. Blocking refers to a first step in a matching algorithm where the number of potential comparisons is limited to rows that are most likely to be a match. The intuition behind blocking is to remove records from consideration with a zero or very low probability of being a match. In our case, we first block on age and estimated income within L2, as it is unlikely that executives in Execucomp data will be under the age of 30 or earn less than $100,000 in income. We focus on a $100,000 threshold instead of considering fine-grained income bins because L2 reports household income generally top coded at $250,000. Moreover, while Excecucomp includes data on yearly compensation and L2 includes household income, the latter values are expected to include the former.[9] As such, restricting our matching procedure to individuals with a reported household income about $100,000 strengthens the probability to obtain correct matches while also serving the purpose of reducing data dimensionality. This restriction leaves 57,786,609 potential matches in the L2 data. The age restriction leaves 43,316,873 potential matches, further reducing the dimensionality of the data and hence the computational complexity and computational resources needed to execute the match. Lastly, we restrict our pool of potential matches to those in the same region as the company's headquarters by blocking on commuting region leaving 38,710,097 potential matches in the L2 data. After blocking on age, income, and commuting region, we implement the Enamorado et al. (2019) algorithm and provide details of this implementation in Appendix B.[10] Next, proceed with processing the raw matched file, as outlined in Figure 1. Specifically, we retain only candidate matches with a posterior probability of above 0.75. We also eliminate all names for which the FastLink algorithm produces multiple candidate matches with a posterior probability above our threshold of choice. Both restricting to matches about the threshold and eliminating duplicates are steps we take to err on the side of caution in interpreting a correct match. However, depending on their research goals, scholars could choose to change the threshold and/or retain and further process duplicates. For example, scholars might choose to include duplicates in regression analyses with weights or collect additional data to discern a correct match among the duplicates.

2.3 | L2-Execucomp matching results

Our matching procedure produced matches for 26,257 executives or 64.5% of the 40,710 executives in the Execucomp data. Table 1 provides summary statistics from all L2 consumer and household data fields for our executives in our matched sample with posterior probabilities ≥95%. For example, we observe that 77% of matched executives buy books, 55% are political contributors, 69% purchase or subscribe to fitness magazines, and 54% donate to charity. In Figure 2, we include a map of the geographical distribution of our matched executives' interests

[9]In addition, the income estimate provided in L2 is estimated using a combination of expenditure patterns and other sources of reported income when available from financial transactions (e.g., loan applications). Thus, the estimated income in L2 differs from the compensation reported in Execucomp in that the estimated income in L2 in part reflects behavior preferences like frugality.

[10]While the steps we took significantly reduced the computational resources needed to execute the match, the sheer size of the data and remaining dimensionality was still sufficiently large that it required significant computing power; we used Amazon Web Services, which allowed us access to multiple parallel 36-core processor virtualized computers (instances) with 160 GB of memory, to execute each attempt to matching in roughly 25 hr per run.

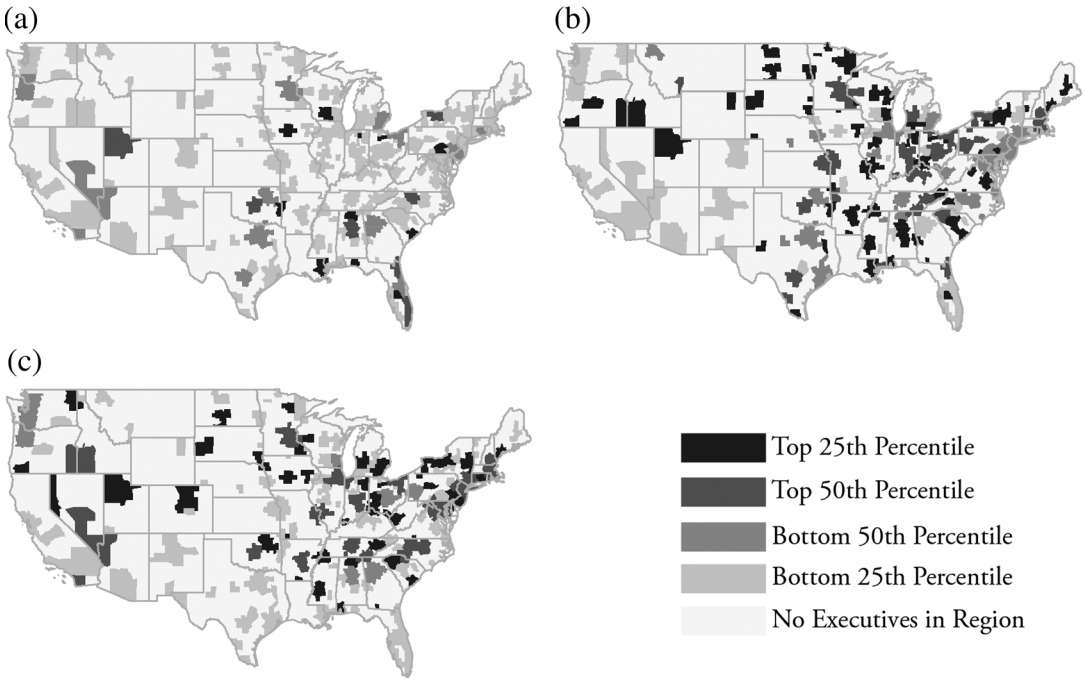

FIGURE 2 Geographical distribution of executive interests in (a) scuba-diving, (b) shooting, and (c) snow skiing

for a subset of our variables (scuba-diving, shooting, and snow skiing), demonstrating that the intensity of interests varies across geographical locations. Some patterns are predicable, such as a higher interest in scuba-diving for executives located near a body of water and a higher interest in shooting in red-leaning regions, while other patterns suggest that executive interests may not correlate strongly with interests we would expect to be prevalent in particular location. This is not necessarily surprising, as executive decisions to live in a particular region are likely to be strongly influenced by public company headquarters and the high income of executives also means that executives can travel to access opportunities elsewhere (e.g., Arizona executives may travel to ski).

As we describe above, one of the benefits of our statistical record-linking approach is that the statistical model used to create matches also provides a continuous probability score reflecting the posterior probability that a given match is correct. Table 2 provides the distribution of matches by the posterior probability of a correct match. As demonstrated in Table 2, approximately 89% of our 26,257 matches (23,548 matches) are estimated to have a posterior probability reflecting at least a 95% chance of being a correct match, and approximately 94% of our 26,257 matches (24,727 matches) are estimated to have a posterior probability reflecting at least a 90% chance of being a correct match. The differences in posterior probability reflect not only partial matches on available data but also the algorithmic process that takes into account the frequency and commonness of variables used in the matching process. Some examples from our matched records illustrate this point. Nolan D. Archibald matched on first name, middle initial, last name, gender, and birthdate with a posterior matching probability of 0.9996. Kevin S. Wilson matched on the exact same attributes as Nolan D. Archibald but received a posterior probability of 0.8088 because the names Kevin and Wilson are more frequent than Nolan and

TABLE 2 Distribution of posterior probability of correct matches between L2 and Execucomp

Posterior probability range	Number of matches	Percentage of matches (%)	Percentage of Execucomp sample (%)	Cumulative percentage of Execucomp sample (%)
95–100%	23,548	89.68	57.84	57.84
90–95%	1,179	4.49	2.90	60.74
85–90%	453	1.73	1.11	61.85
80–85%	626	2.38	1.54	63.39
75–80%	451	1.72	1.11	64.50
Total	26,257	100	64.50	

Archibald.[11] Craig Brown, another individual with a first and last name that occur more frequently received a posterior probability of 0.76287 because, in addition to the frequency of his first and last names, the matching algorithm retrieved only a partial match on birthdate, that is, a difference greater than 1 year but shorter than 3 years.

The resulting crosswalk between L2 and Execucomp contains a column with the continuous posterior probability associated with each match, thereby enabling researchers who use our crosswalk to choose the probability cutoff that best fits their purposes and/or to test the sensitivity of their results.

2.3.1 | L2-Execucomp matching benchmarking

Although the high posterior probabilities reported in Table 2 suggest that our matching procedure using the Enamorado et al. (2019) algorithm generates a substantial number of high-quality matches, we take additional steps to ensure the accuracy and reliability of these results. Specifically, we benchmarked our matching approach against prevailing record-linking techniques using string matching on names *and* employer as a discriminating variable. Because employer is not available in the L2 data, we turn to a new database to benchmark our results, the Database on Ideology, Money in Politics, and Elections (DIME)[12] (Bonica, 2014). The advantage of the DIME database is that it not only includes employer which can be used as a discriminating variable to match to Execucomp using prevailing record-linking techniques but also includes home address which can be used to verify if the resulting L2-Execucomp-DIME match is accurate.

First, the DIME database includes contributor first name, last name, and self-reported employer, thereby allowing us to match the DIME data to Execucomp following prevailing record-linking techniques which uses employer as a discriminating variable in conjunction with first and last names to generate matches (e.g., Chin & Semadeni, 2017; Gupta, Briscoe, & Hambrick, 2018). Next, because we matched the L2 data to Execucomp and we matched

[11]Note that name frequencies and other match criteria are calculated within the commuting region of the executive. Thus, two matched executives with the same name may receive different posterior probabilities if they are matched in different commuting regions where the commonness of their first and last name varies between these regions.

[12]The Database on Ideology, Money in Politics, and Elections (DIME) database aggregates political contribution data and disambiguates contributor names, assigning each a stable and unique identifier (Bonica, 2013, 2014).

Execucomp to the DIME data, we can isolate executives for whom we are able to match in all datasets. That is, executives for whom we were able to match to the L2 data using our matching strategy *and* to the DIME using prevailing methods. This allows us to benchmark the accuracy of the L2 to Execucomp match against the classic matching approach employed in the case of Execucomp to DIME, by exploiting the fact that both the L2 data and the DIME data report individual home addresses. The intuition behind this exercise is that if our matching procedure using the Enamorado et al. (2019) method performed well, then, for an L2-Excecomp-DIME matched record, the home address reported in the L2 data and the home address reported in the DIME should overlap. A high degree of overlap would suggest that our matching procedure performs well relative to existing string-based matching techniques, which are possible only when quality discriminating variables such as employer are available.[13] Given the benchmarking purpose of this exercise, we match Execucomp to the DIME data using an exact match criterion for first name, last name, and employer—the most stringent implementation of string matching. While this limits the number of total matches between the Execucomp and DIME data, it also minimizes the rate of false positives, ensuring that our Execucomp to DIME crosswalk could be used as "ground truth" to assess the quality of the L2-Execucomp matches.

We start with the set of 34,795,868 DIME contribution records from 2000 to 2014 and restrict to the subset of 21,946,967 contribution records that provided information on employer. This represents 63% of the 2000–2014 DIME data. Next, we exact match the DIME records to Execucomp using first name, last name, and employer and obtain 27,833 matched contribution records. Next, we further restrict the matched dataset to contribution records that report a home address. We manually check the reported addresses and observe that 12% of the matched records have incomplete addresses (e.g., only a city is mentioned), 11.5% report an office address, 4% report a P.O. Box, and the remaining report a home address. We manually verify the distinction between home and office address by coding an address as an office address when the contributor's reported home address corresponds to his/her reported employer headquarters or when the information provided is a commercial address. Last, we deduplicate the dataset to remove repeat contribution records for the same executive unique identifier. We take the most recent home address record for each executive. This results in a dataset of 4,026 exact matches of executive unique identifiers that report a home address in DIME between 2000 and 2014 which represents 15.33% of our sample.

In Figure 3, we provide a histogram of the observed geographical distance between the address in the L2 data and the self-reported home address in the DIME data for matched executives using latitude–longitude coordinates. In Figure 3a, we use all the data, including contributions from 2000 to 2014. In Figure 3b, we restrict to data where contributions are made between 2012 and 2014, leaving 1,446 contributions. In the longer run dataset, individuals have more time to move from their reported home address. Thus, we expect to observe records with a larger estimated distance in the 2000–2014 dataset than in the 2012–2014 dataset. Indeed, we observe that the maximum distance in Figure 3a is 468.5 miles, whereas in Figure 3b it is only 174.8 miles. Nevertheless, in both cases, we observe that the vast majority of matched individuals have a distance less than 10 miles away between the L2 and DIME addresses. Specifically,

[13]One might be concerned that the benchmarking exercise is biased to the extent that there is a correlation between commonness of names and propensity to report an address in DIME. The correlations between the propensity to report an address in DIME and (a) the frequency of last names in the U.S. population and (b) the frequency of first names in the U.S. population are (a) 0.011 and (b) 0.035, respectively. Similarly, the correlations between the geographic distance between the addresses in DIME and those in L2 and (a) the frequency of last names in the U.S. population and (b) the frequency of first names in the U.S. population are (a) 0.007 and (b) 0.026, respectively.

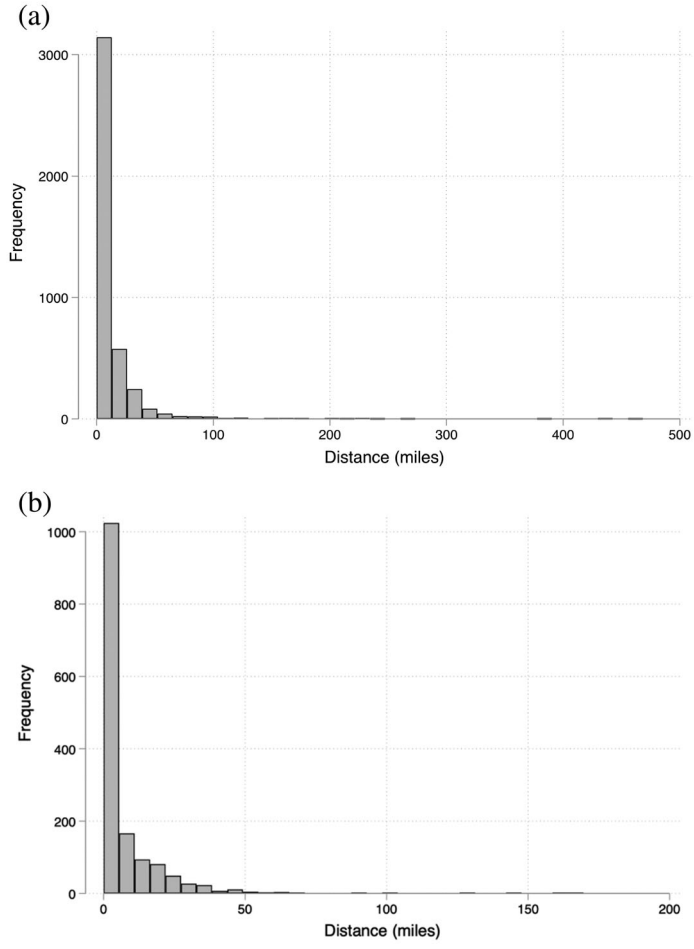

FIGURE 3 Frequency plot of geographic distance between home addresses reported in L2 and Database on Ideology, Money in Politics, and Elections (DIME) for executives in Execucomp matched to both data sources. (a) 2000–2014 and (b) 2012–2014

in Figure 3a, 72.6% of individuals have a distance of less than 10 miles away and 96.1% of individuals have a distance of less than 50 miles away. In Figure 3b, 72.6% of individuals have a distance of less than 10 miles away and 96.4% of individuals have a distance of less than 50 miles away. Thus, we interpret Figure 3 as providing credible evidence that despite the significant challenges associated with matching the L2 data to Execucomp, our matching strategy provides quality matches consistent with expectations of prevailing record-linking techniques.

There might be a concern that there is a systematic correlation between the type of individual who selects into reporting (or not reporting) employer information in DIME and the type of names that are more prone to being (or not being) matched by our L2 to Execucomp matching algorithm. If that were the case, then the benchmarking exercise we describe in this section would be biased. To check this possibility, we tested the correlation value between frequency of names and propensity of those names to report an employer in DIME. Specifically, the accuracy of the matching algorithm we use, and hence the probability to obtain a match, decreases with the frequency of the name in the population. In other words, for frequent names

such as Smith, the algorithm detects many possible matches, hence labeling any potential match as having a low accuracy probability. It follows that if there is any bias in our validation exercise, it has to do with a correlation between names that are frequent in the U.S. population and names that chose to not report the employer information in DIME. We calculate the frequency with which each name in DIME reports the employer information as the fraction of names of each type that reports this information. We then calculate the correlation between this value and the number of times the name is present in the U.S. population. We obtain a correlation of −0.002, which indicates that there is no relationship between the two.

3 | DISCUSSION AND FUTURE RESEARCH

In this article, we introduce a new dataset that provides a novel look into the revealed preferences and off-the-job interests of public firm top executives as reflected in household credit card spending patterns. To do so, we matched data from the commercial data provider L2 which sources credit card spending data from the credit reporting agency Experian with executives listed in Execucomp. To overcome the computational and technical challenges associated with the linking of these two data sources, we leveraged insights from the statistical record-linking literature and implemented a matching strategy which allowed us to link records across the two data sources despite limited overlapping information. Our final database contains 94 revealed preference indicators of off-the-job interests as reflected by credit card spending patterns matched to 26,257 top executives. We make our crosswalk freely available for academic use, the goal being to facilitate and stimulate research that explores the multitude of new research questions made possible with these data.

By making our matched crosswalk publicly available for academic use, we see a number of avenues for future research and new research opportunities. As a start, researchers may wish to begin with discovery-driven machine learning exercises to explore patterns which emerge between the revealed preferences of top executives and a variety of outcomes of interest to strategy scholars (Choudhury, Allen, & Endres, 2021). The purpose of such exercises would not be to deductively test theory, but rather to use a data-driven approach for pattern discovery which could serve as the basis for future inductive and abductive theorizing. While some revealed preference variables likely represent proxies for existing conceptual concepts with established theoretical linkages, other variables that are demonstrated to have predictive validity may inform new ex post theorizing, construct development, and ultimately the development of new and rich conceptual linkages. At the same time, many of the revealed preference variables in our data are thematically linked and hence may be reflectors of higher order theoretical constructs. For instance, researchers have used lists of "risk" hobbies from actuaries to develop measures of executive risk-taking propensities (Kim & Kamiya, 2015). While some scholars may wish to use different variables in our dataset independently, others may wish to combine individual indicators in a way that capture higher order constructs—either in a theory-driven deductive manner or by an empirical approach such as factor analysis.

While data-driven and machine learning techniques are useful to uncover patterns in the data, provide insights which guide new theory development, and are likely to become increasingly important as the flood of "big data" advances upon management researchers, the data we construct also provide an opportunity for researchers to conveniently measure a number of potentially difficult-to-measure theoretical constructs at a large scale, thereby facilitating deductive theory testing and catalyzing future research by minimizing the investments needed to

compile new data sources. For example, Sunder et al. (2017) hand-collect data from the Federal Aviation Administration to identify pilot CEOs, arguing that being a pilot reflects "sensation-seeking" which they find to be related to innovative firm outcomes. Our data provide a trove of information on executives which should enable scholars to build on this line of work in an easy to use and large-scale way. For instance, sensation-seeking may be proxied by executive interests in action sports such as auto racing and motorcycling and could be compared and contrasted with proxies for risk-taking behavior, such as interests in gambling and gaming casinos. Likewise, Campbell and Zipay (2019) argue that the types of leisure activities CEOs engage in—hand-collecting data on CEO marathon running—will be associated with shirking behavior. Our comprehensive array of executive hobbies and interests would allow for a deeper exploration between the leisure actives and firm outcomes in an easy to use and large-scale format (Campbell & Zipay, 2019).

Conceptually, top executives who have atypical interests—interests in a wider variety of hobbies and activities—may manifest these atypical preferences in firm diversification decisions. It is plausible that executive interests may also influence classic make or buy decisions. While transaction cost economics dictates that vertical integration is influenced by factors such as asset specificity, there is growing evidence that such decisions may be influenced by executive characteristics and cognition (Foss & Weber, 2016; Weber & Mayer, 2014). Such decisions may also be influenced by executive preferences. For example, an interest in do-it-yourself, home repair, auto repair, and other related factors may predict a firm's decision to build versus buy or moderate the magnitude of relationships predicted by transaction cost theory. Personal interests in international travel and cuisines may also be associated with a firm's decision to expand operations internationally. Personal interests in family, exercise, health, and nutrition may be a key explanation as to why firms adopt various work practices (Bloom, Kretschmer, & Van Reenen, 2011). Interests in the arts, theater, and other artistic activities may result in CEOs allocating more resources to advertising campaigns. Our data allow for testing these and other similar questions at scale.

Our dataset also allows scholars to explore a number of questions related to TMT dynamics (Neely Jr et al., 2020). Because our resulting dataset matches revealed preferences to a firm's top executives, not just the CEO, our data provide a unique opportunity to explore issues such as cultural matching in the upper echelons (Rivera, 2012), to what degree TMTs are comprised of executives with homogeneous off-the-job interests, and what, if any, impact such compositions may have on organizational outcomes. For example, researcher may make use of prior work which has identified sudden and unexpected CEO deaths (Quigley, Crossland, & Campbell, 2017) as a way to explore whether replacement CEOs reflect relative homogeneity with respect to off-the-job interests of the existing TMT. To that end, scholars may leverage our matching code and statistical matching approach to link the L2 data to databases reflecting corporate boards of directors. Doing so would allow researchers to extend the analysis of off-the-job interests to corporate boards and study how such interests may correlate with board formation and board interlocks (Howard, Withers, & Tihanyi, 2017), the relationship between board and TMTs (Westphal, 1999), or other outcomes such as corporate misconduct (Neville, Byron, Post, & Ward, 2019). Likewise, researchers could make use of the recently developed database on CEO turnover and CEO dismissals (Gentry, Harrison, Quigley, & Boivie, 2021) to explore how CEO interests, the composition of TMT interests within the firm, or the relationship between off-the-job interests of CEOs and the corporate board may predict CEO turnover or dismissal, or how such interests may moderate the intensity of established relationships such as the relation between poor firm performance and the likelihood of CEO dismissal.

Lastly, our data provide detailed information on a new dimension of executive characteristics which may contribute to work regarding social and behavioral influences on CEOs, TMTs, and the CEO-TMT or CEO-Board interface. As recently described by Bromiley and Rau (2016) in a survey of the literature, existing work has focused on three approaches to study such issues: observable characteristics, unobservable characteristics related to personality or other psychological constructs, and characteristics of the broader group such as social ties. Our article suggests the potential salience of a fourth approach, off-the-job interests of top executives, and provides the corresponding data which would facilitate the empirical exploration of this approach at a large scale.

4 | CONCLUSION

In this article, we introduce a new database that provides a heretofore unprecedented look into the personal lives and off-the-job interests of U.S. public firm executives. To construct this database, we match novel household credit card spending data sourced from credit reporting agency Experian and provided by the commercial data provider L2 with the population of top executives listed in Execucomp. To overcome the computational challenges associated with matching these data, we build on the statistical record-linking literature in a way that allows us to generate reliable matches with limited information. Our matched database provides information pertaining to 94 off-the-job interests for over 26,000 top executives, as determined by household credit card spending patterns. We make our matching crosswalk and corresponding code freely available for academic use, the goal being to create a public good facilitating future research.

ACKNOWLEDGEMENTS

The authors thank the editor and two anonymous referees for their helpful comments and suggestions. They also thank the research assistants who assisted with this project.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from L2 and Execucomp. Restrictions apply to the availability of these data, which were used under license for this study. All of our matching code is available on our Github repository, as will be the resulting matched crosswalk between L2 and Execucomp identifiers. This crosswalk will allow researchers to leverage our substantial matching efforts and easily integrate the two sources with the relevant licenses.

REFERENCES

Biggerstaff, L., Cicero, D. C., & Puckett, A. (2017). FORE! An analysis of CEO shirking. *Management Science*, *63*(7), 2302–2322.

Bloom, N., Kretschmer, T., & Van Reenen, J. (2011). Are family-friendly workplace practices a valuable firm resource? *Strategic Management Journal*, *32*(4), 343–367.

Bonica, A. (2013). Ideology and interests in the political marketplace. *American Journal of Political Science*, *57*(2), 294–311.

Bonica, A. (2014). Mapping the ideological marketplace. *American Journal of Political Science*, *58*(2), 367–386.

Bromiley, P., & Rau, D. (2016). Social, behavioral, and cognitive influences on upper echelons during strategy process: A literature review. *Journal of Management*, *42*(1), 174–202.

Buss, D. M. (1985). Human mate selection: Opposites are sometimes said to attract, but in fact we are likely to marry someone who is similar to us in almost every variable. *American Scientist*, *73*(1), 47–51.

Cain, M. D., & McKeon, S. B. (2016). CEO personal risk-taking and corporate policies. *Journal of Financial and Quantitative Analysis*, *51*(1), 139–164.

Campbell, R. J., & Zipay, K. (2019). Not all leisure is shirking: CEO endurance leisure and firm value. *Academy of Management Proceedings*, *2019*(1), 12155.

Cannella, B., Finkelstein, S., & Hambrick, D. C. (2008). *Strategic leadership: Theory and research on executives, top management teams, and boards*. New York, NY: Oxford University Press.

Carpenter, M. A., Geletkanycz, M. A., & Sanders, W. G. (2004). Upper echelons research revisited: Antecedents, elements, and consequences of top management team composition. *Journal of Management*, *30*(6), 749–778.

Chin, M., Hambrick, D. C., & Treviño, L. K. (2013). Political ideologies of CEOs: The influence of executives' values on corporate social responsibility. *Administrative Science Quarterly*, *58*(2), 197–232.

Chin, M., & Semadeni, M. (2017). CEO political ideologies and pay egalitarianism within top management teams. *Strategic Management Journal*, *38*(8), 1608–1625.

Choudhury, P., Allen, R. T., & Endres, M. G. (2021). Machine learning for pattern discovery in management research. *Strategic Management Journal*, *42*(1), 30–57.

Choudhury, P., Wang, D., Carlson, N. A., & Khanna, T. (2019). Machine learning approaches to facial and text analysis: Discovering CEO oral communication styles. *Strategic Management Journal*, *40*(11), 1705–1732.

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Berlin, Heidelberg: Springer.

Christensen, D. M., Dhaliwal, D. S., Boivie, S., & Graffin, S. D. (2015). Top management conservatism and corporate risk strategies: Evidence from managers' personal political orientation and corporate tax avoidance. *Strategic Management Journal*, *36*(12), 1918–1938.

Cronqvist, H., Makhija, A. K., & Yonker, S. E. (2012). Behavioral consistency in corporate finance: CEO personal and corporate leverage. *Journal of Financial Economics*, *103*(1), 20–40.

Davidson, R., Dey, A., & Smith, A. (2015). Executives' "off-the-job" behavior, corporate culture, and financial reporting risk. *Journal of Financial Economics*, *117*(1), 5–28.

Enamorado, T., Fifield, B., & Imai, K. (2019). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, *113*(2), 353–371.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210.

Foss, N. J., & Weber, L. (2016). Moving opportunism to the back seat: Bounded rationality, costly conflict, and hierarchical forms. *Academy of Management Review*, *41*(1), 61–79.

Gamache, D. L., Neville, F., Bundy, J., & Short, C. E. (2020). Serving differently: CEO regulatory focus and firm stakeholder strategy. *Strategic Management Journal*, *41*(7), 1305–1335.

Gentry, R. J., Harrison, J. S., Quigley, T. J., & Boivie, S. (2021). A database of CEO turnover and dismissal in S&P 1500 firms, 2000–2018. *Strategic Management Journal*, *42*(5), 968–991.

Graffin, S. D., Hubbard, T. D., Christensen, D. M., & Lee, E. Y. (2020). The influence of CEO risk tolerance on initial pay packages. *Strategic Management Journal*, *41*(4), 788–811.

Gupta, A., Briscoe, F., & Hambrick, D. (2018). Evenhandedness in resource allocation: Its relationship with CEO ideology, organizational discretion, and firm performance. *Academy of Management Journal*, *65*(5), 1848–1868.

Hambrick, D. C. (2007). Upper echelons theory: An update. *Academy of Management Review*, *32*(2), 334–343.

Hambrick, D. C., & Mason, P. (1984). Upper echelons: The organization as a reflection of its top managers. *Academy of Management Review*, *9*(2), 193–206.

Harrison, J. S., Thurgood, G. R., Boivie, S., & Pfarrer, M. D. (2019). Measuring CEO personality: Developing, validating, and testing a linguistic tool. *Strategic Management Journal*, *40*(8), 1316–1330.

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. New York, NY: Springer Science & Business Media.

Hill, A. D., Recendes, T., & Ridge, J. W. (2019). Second-order effects of CEO characteristics: How rivals' perceptions of CEOs as submissive and provocative precipitate competitive attacks. *Strategic Management Journal*, *40*(5), 809–835.

Howard, M. D., Withers, M. C., & Tihanyi, L. (2017). Knowledge dependence and the formation of director interlocks. *Academy of Management Journal*, *60*(5), 1986–2013.

Kim, Y. & Kamiya, S. (2015). The testosterone of the CEO and the risk of the firm. (January 29, 2015).

Lawrence, B. S. (1997). Perspective—The black box of organizational demography. *Organization Science*, *8*(1), 1–22.

Limbach, P., & Sonnenburg, F. (2015). *Does CEO fitness matter? CFR Working Paper*.

Neely, B. H., Jr., Lovelace, J. B., Cowen, A. P., & Hiller, N. J. (2020). Metacritiques of upper echelons theory: Verdicts and recommendations for future research. *Journal of Management*, *46*(6), 1029–1062.

Neville, F., Byron, K., Post, C., & Ward, A. (2019). Board independence and corporate misconduct: A cross-national meta-analysis. *Journal of Management*, *45*(6), 2538–2569.

Ouyang, B., Tang, Y., Wang, C., & Zhou, J. (2021). No-fly zone in the loan office: How chief executive Officers' risky hobbies affect credit Stakeholders' evaluation of firms. *Organization Science*, *33*, 414–430.

Quigley, T. J., Crossland, C., & Campbell, R. J. (2017). Shareholder perceptions of the changing impact of CEOs: Market reactions to unexpected CEO deaths, 1950–2009. *Strategic Management Journal*, *38*(4), 939–949.

Rivera, L. A. (2012). Hiring as cultural matching: The case of elite professional service firms. *American Sociological Review*, *77*(6), 999–1022.

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, *5*(17), 61–71.

Sunder, J., Sunder, S. V., & Zhang, J. (2017). Pilot CEOs and corporate innovation. *Journal of Financial Economics*, *123*(1), 209–224.

Wang, G., Holmes, R. M., Jr., Oh, I. S., & Zhu, W. (2016). Do CEOs matter to firm strategic actions and firm performance? A meta-analytic investigation based on upper echelons theory. *Personnel Psychology*, *69*(4), 775–862.

Wasi, N., & Flaaen, A. (2015). Record linkage using Stata: Preprocessing, linking, and reviewing utilities. *The Stata Journal*, *15*(3), 672–697.

Weber, L., & Mayer, K. J. (2014). Transaction cost economics and the cognitive perspective: Investigating the sources and governance of interpretive uncertainty. *Academy of Management Review*, *39*(3), 344–363.

Westphal, J. D. (1999). Collaboration in the boardroom: Behavioral and performance consequences of CEO-board social ties. *Academy of Management Journal*, *42*(1), 7–24.

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods. American Statistical Association*, 354–359.

Wowak, A. J., Mannor, M. J., Arrfelt, M., & McNamara, G. (2016). Earthquake or glacier? How CEO charisma manifests in firm strategy over time. *Strategic Management Journal*, *37*(3), 586–603.

How to cite this article: Raffiee, J., Fehder, D., & Teodoridis, F. (2022). Revealing the revealed preferences of public firm CEOs and top executives: A new database from credit card spending. *Strategic Management Journal*, *43*(10), 2042–2065. https://doi.org/10.1002/smj.3397

APPENDIX A

A.1 | Additional L2 fields

Table 1 includes data on revealed preferences for individuals and households. L2 provides a number of other fields that could be used in research projects. The dataset includes political variables and demographic information about the household in which the individuals live. Below, we provide a reference to these variables. In Table A1, we list the political variables. In Table A2, we provide a reference to the household demographic information.

TABLE A1 Political variables

Variable	Observation level	Short description
Election returns	State, county, and precinct	Election results for local, state-wide, and national elections at various levels of aggregation
Voter turnout	State, county, and precinct	Share of registered voters who voted in local, state-wide, and national elections
Vote frequency	Individual	Information about whether individual voted in local, state-wide, and national elections
Voter registration	Individual	Party affiliation of the individual (if disclosed by state)
Absentee or early voting	Individual	Whether individual is registered to vote through absentee or early voting (if disclosed by state)

TABLE A2 Household information

Variable	Short description
Household size	Number of individuals in household
Household gender	Gender mix of household
Presence of children	Whether children are in household
Household ethnic composition	Inferred ethnic composition of household
Spanish speaking	Spanish speaking individuals in household
Home owner/renter	Does household rent or own dwelling
Home purchase price	Purchase price of home if owned
Home purchase date	Date of the home's last purchase
Estimated home value	Inferred value of home
Dwelling type	Whether dwelling is a free-standing home, apartment, and so on
Land value	Estimated unimproved value of land in which the household's dwelling sits
County ethnic composition	Census estimates of the ethnic composition of the county in which the household's dwelling is contained
County median income	Census estimates of median income for county in which the household's dwelling is contained
County median education years	Census estimates of median number of years of education for residents of county in which the household's dwelling is contained

APPENDIX B

B.1 | Implementing the Enamorado et al. algorithm

To match the L2 data to the Execucomp data, we make use of the algorithm described in Enamorado, Fifield, and Imai (2019). To do so, we build off of their open-source code available on a public Github repository.[14] While their publicly available code takes care of many of the details for implementing the Fellegi and Sunter (1969) record linkage methodology using the expectation maximization algorithm, there were still a number of choices, built within the code, that had to be made to fit our particular empirical setting. We describe these in the following.

We make our resulting code applying the Enamorado et al. (2019) algorithm publicly available via a Github repository.[15] The code repository is meant to enable researchers to execute our code directly in their chosen environment. In this appendix, we provide a description of the different files in our repository and steps our code takes to achieve the matching.

B.2 | Commuting region definitions

To explain our coding of Commuting Region, we begin by explaining the U.S. Census's definitions of Core-Based Statistical Areas. Each county in the United States is classified by the Census as either a metropolitan statistical area (50,0000 or more population) (MSA), a micropolitan statistical area (at least 10,000 population but more than 50,000) or rural. The Census then takes each MSA and adds all counties that have a substantial economic relationship with the core MSA as defined as having at least 25% of workers commuting between the county and the urban core. The final MSA is a list of counties that have substantial economic and social ties with a core urban area of at least 50,000. The Census Bureau also defines combined statistical areas (CSAs). This definition begins in the same way as an MSA but adds counties where at least 15% but less than 25% of the population of the county commutes to the core urban area. All CSAs have a subset of the region that are also an MSA but not all MSAs are part of a CSA. Every county, MSA and CSA have a Census defined region code and these codes essentially form a hierarchy starting at the county level and potentially flowing up to an MSA code and then potentially to a CSA code.

Building on this logic, we begin our commuting region definition at the CSA level. For each county in the United States, we assign it the Commuting Region Code of the CSA of which it is a member, if it is a member of the CSA. If the county is a member of an MSA but not a CSA, we give the county the Commuting Region Code of the MSA. If the county is a member of a micropolitan area but not an MSA or CSA, we give that county the Commuting Region Code of the Micropolitan Area Code. Lastly, if a county is not a member of a CSA, MSA or micropolitan area, we give it the Commuting Region Code of the county code.

Using this coding scheme, we are able to code each company headquarters address in Execucomp and each residential address in the L2 data with a Commuting Region Code. We consider residential addresses in L2 as a match to the company headquarters address in Execucomp if they share the same region code.

[14] https://github.com/kosukeimai/fastLink.

[15] https://github.com/dfehder/FRT_2022_SMJ_CEO_TopExec_Preferences.

B.3 | Data preprocessing

Before the data can be put through the Enamorado et al. (2019) algorithm, it needs to be processed from its original raw form to fit the algorithms' required format. We achieved these processing steps through a set of two Python program files, one for each of the two datasets: preprocess_execucomp.job for the Execucomp data (in the preprocess_execucomp folder) and l2-0-master.job for the L2 data (in the preprocess_l2 folder). The details of the preprocessing steps are well-documented in the code as well as in the accompanying README file in the Github repository.

B.4 | Executing the Enamorado et al. algorithm

The execution of the matching algorithm is achieved through two R program files. At the highest level, run_fastlink.job is a script that loads the data into the Enamorado et al. (2019) algorithm and then processes and outputs the results. The core matching algorithm is contained in runFastlink.R. This file is drawn directly from the Github algorithm provided by Enamorado et al. (2019).

B.5 | Final crosswalk production

The FastLink algorithm returns produces multiple candidate matches with a posterior probability above our threshold of choice (0.75). We first aggregate all matches from each commuting region into one file. The next step of our final crosswalk production is to refine the matches so that the best possible match is selected for each executive in Execucomp as measured by posterior probability. We also eliminate all names for which the FastLink algorithm produces multiple candidate matches with a posterior probability above our threshold of choice. The script that executes each of these steps is contained in agg_match.job.