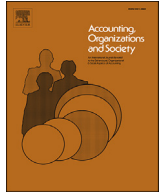Contents lists available at ScienceDirect

Accounting, Organizations and Society

journal homepage: www.elsevier.com/locate/aos

Commentary

The promise and challenges of new datasets for accounting research☆

Siew Hong Teoh

Paul Merage School of Business, University of California, Irvine, United States

ARTICLE INFO

ABSTRACT

I describe a brief summary of the development of databases used in accounting research and discuss the research questions addressed in traditional databases and 'new' databases. The new data include online searches such as Google Trends data; textual data from corporate disclosures, analyst reports, conference call transcripts, earnings press releases, and news media articles; social network and social media data from Twitter, LinkedIn, Glassdoor, and other data. New data holds promise for research on attention or cognitive processing constraints; on tone/valence, affect, deceptiveness and credibility for capital market and financial reporting outcomes. I examine the econometric challenges of new data and suggest the potential for new data to offer new auditing tools to detect poor financial reporting, which will help to discourage earnings management.

1. Introduction: accounting data as the quantitative expression of economic phenomena

Davidson (1966) describes accounting as the study of the quantitative expression of economic phenomena. Over the last half century there has been an enormous blossoming of research in accounting—especially the capital markets area—made possible by machine-readable datasets (beginning with CRSP and Compustat) along with improvements in data and statistical technology for processing and analyzing the data (WRDS).

In the last 25 years, at a rapidly accelerating pace, there has been an explosive growth in new types of data, or traditional data that have become newly machine-readable and analyzable, that can provide insight into economic phenomena. These include data on online search behaviors; the texts of corporate disclosure, annual reports, conference calls, managerial forecasts, and analyst reports; data on the structure of social networks and transmission of information along the networks about managers, directors, analysts, and investors; and data that is neither verbal nor numerical, such as tone of voice and images.

The new datasets being generated in many areas from the online revolution are often called `big data.' Of course, the major traditional datasets used in accounting research are also large, and are therefore `big data.' But the new data sources provide distinctive types of information that allow addressing a wider range of issues effectively. Indeed, new forms of data are leading to major changes throughout the social sciences. The interdisciplinary field of social networks is booming in economics, computer science, sociology, and many other fields, because of both advances in theory and the rise of social networking technologies and applications; see the surveys of Jackson (2010) and Jackson, Rogers, and Zenou (2017). It should be no surprise that data relating to online behavior and

social networks is important for accounting research as well.

There are also tremendous innovations in the ability of researchers to generate new data targeted specifically to answer research questions. Modern software has made it more efficient to run experiments both in the experimental laboratory and especially through the internet. Amazon Mechanical Turk makes it feasible to recruit very large numbers of subjects at very low cost—subjects who are potentially more representative of a wider population than the college students typically drawn upon for experiments by academic scholars (see Farrell, Grenier, & Leiby, 2017; Rennekamp 2012). Software improvements and a growing ethos of scientific experimentation by firms has also led to a great increase in A-B testing (tests of alternative business procedures) by firms. This has created an opening for scholars (so far more in economics and finance than in accounting) to generate new data by running experiments in the field (but in accounting, see Lawrence, Ryan, Sun, & Laptev, 2018). Furthermore, the field of neuroeconomics has grown along with improvements in the ability to measure brain activity (especially, fMRI technology; in accounting see Farrell, Goh, & White, 2014) and various other forms of physiological activity including skin conductance and hormonal responses.

In this article, I discuss the general promise and challenges of these new datasets for accounting research. My thesis is that this explosion of new data promises to lead to a revolutionary improvement in our understanding of accounting phenomena. I explain here the reasons for this optimism, and give some tentative suggestions for further directions for exploiting new datasets that are available now, and those that will foreseeably be available soon. By way of example, I discuss how the paper "A New Era of Voluntary Disclosure? Empirical Evidence on How Employee Postings on Social Media Relate to Future Corporate Disclosures" by Hales et al. (2018) fits into the larger scope of new research questions that can be studied with new data, as well as the limitations and challenges research with new data will likely face. Existing research using the new generation of datasets has already provided valuable new insights, and as the new data and techniques for analyzing them continue to improve, these approaches will allow us to provide better answers to important questions that, so far, have been hard to address.

For example, popular observers have long claimed that capital markets behavior is influenced by investor moods, leading to manias and panics. It seems intuitively plausible that an investor who is excited will react to new accounting information differently from one who is terrified or suspicious of the news, and that at the aggregate level, this might lead to market booms or crashes. But such claims are hard to evaluate rigorously. Using just traditional data—stock prices, trades, and accounting numbers—it is hard to pin down the role of feelings in capital markets. As I will discuss, modern textual analysis of some crowd sourced reviews can provide measurements of expressed emotions to test such ideas, and more importantly to develop more refined and testable theories. Ultimately, it will also permit testing of whether accounting acts to discipline waves of emotion in capital markets, or whether, perhaps, under some circumstances accounting disclosures exacerbate the problem. Similarly, I will argue that new datasets permit a much more refined empirical study of attention allocation by investors. This makes it feasible to address a variety of new questions about investor attention to accounting information and resulting effects on market prices.

To understand the promise of new datasets, it is useful to review how the appearance of new datasets in the past promoted scientific accounting research. I do this in the next section. In Section 3, I consider systematically what types of research questions researchers may be able to address with different databases—both traditional and new. In addition to standard issues such as value

relevance of different kinds of information, and market efficiency in processing public information, I argue that new datasets are especially useful in providing insight about the allocation of attention by managers, intermediaries and investors; the effects of feelings in markets; and how social interactions affect economic behavior. In Section 4, I consider econometric challenges in the use of new datasets. Finally, in Section 5, I conclude and suggests directions for the use of new datasets in future research.

2. Brief history of development of accounting databases and scientific research

The availability of machine-readable datasets on securities trading and prices from the Center of Security Prices at the University of Chicago in 1960 and financial data items from Compustat in 1962 fueled the growth in capital markets research in accounting. The availability of datasets on analyst earnings estimates quickly followed around 1976-77 from the Institutional Brokers' Estimate System (I/B/E/S) and Zacks.

In 1993, Wharton Research Data Services (WRDS) was developed, which increased the efficiency of empirical researchers in analyzing a broad range of data. Prior to WRDS, each individual researcher would spend considerable time writing Fortran or SAS codes to query databases to extract information they need, and then reformatting the data into a form suitable for use in statistical analyses. WRDS offered a user-friendly interface for any university subscriber to quickly extract and format data, collated many different databases for a one-stop location for obtaining data, and offered access to database experts previously unavailable to universities with limited resources.[1] In short, easier access to quantifiable data made possible by WRDS and other data aggregators spurred the volume, depth, and scope of academic scientific research in the social sciences in the last 25 years across a broad range of universities, from lower tier to elite universities.

The release of the first popular web browser Mosaic in 1993 ushered in explosive growth in internet usage by the public, and continued to grow exponentially into the 2000s. Online marketplaces were born (Amazon.com, 1994; Alibaba.com, 1999), search directories (Yahoo!, 1995) and search engines (Google, 1998; Baidu, 2003) for online information came next, and social networking sites proliferated (AOL instant messenger for internet chats, 1997; MySpace, 2002; LinkedIn, 2003; Facebook, 2004; Yelp, 2004; YouTube, 2005; Twitter, 2006; Glassdoor, 2007). Each of these websites offer researchers with new sources of data about transactions, crowd sourced opinions, and social interactions.

3. What types of research issues do different databases allow us to address?

3.1. Traditional databases

A vast array of databases has been used in financial accounting research, but especially dominant are databases with numerical information from corporate accounting reports and disclosures, security prices and trades, and analyst forecasts and recommendations. I will refer to these as "traditional databases."

Exploration of a wide range of research questions has been

---

[1] Since its founding, WRDS has grown tremendously in terms of subscribers to "over 50,000 commercial, academic, and government users in 30 + countries" and in terms of the number of databases it covers to "over 250 terabytes of data across multiple disciplines including Accounting, Banking, Economics, ESG, Finance, Healthcare, Insurance, Marketing, and Statistics" (http://www.whartonwrds.com/about/).