Journal of Supply Chain Management
2020, 56(2), 7–18

THEORIZING SUPPLY CHAINS WITH QUALITATIVE BIG DATA AND TOPIC MODELING

PRATIMA (TIMA) BANSAL, JURY GUALANDRIS, AND NAHYUN KIM
Western University

The availability of Big Data has opened up opportunities to study supply chains. Whereas most scholars look to *quantitative* Big Data to build theoretical insights, in this paper we illustrate the value of *qualitative* Big Data. We begin by describing the nature and properties of qualitative Big Data. Then, we explain how one specific method, topic modeling, is particularly useful in theorizing supply chains. Topic modeling identifies co-occurring words in qualitative Big Data, which can reveal new constructs that are difficult to see in such volume of data. Analyzing the relationships among constructs or their descriptive content can help to understand and explain how supply chains emerge, function, and adapt over time. As topic modeling has not yet been used to theorize supply chains, we illustrate the use of this method and its relevance for future research by unpacking two papers published in organizational theory journals.

Keywords: *qualitative research; Big Data; topic modeling; complex adaptive systems; networks*

INTRODUCTION

The advent of digital technologies has opened up opportunities for researchers to gather large volumes of data, often called Big Data,[1] to theorize supply chains. Quantitative Big Data can inform supply chain theory because the large volume of granular data can describe all aspects of supply chains, including customer Web site visits, customer sentiments from social media, varying service levels, or evolving contractual ties and material flows in a supply chain (Miller, Ganster & Griffis, 2018; Mishra et al., 2018; Park, Bellamy & Basole, 2018).

Supply chain researchers thus far have largely focused on *quantitative* Big Data, with *qualitative* Big Data being largely overlooked. However, qualitative Big Data can reveal how supply chains emerge, function, and adapt, complementing insights garnered from quantitative Big Data. Qualitative Big Data offers rich, contextualized information about multiple, diverse organizations and their complex connections. Whereas quantitative data often require researchers to predetermine constructs and operationalizations, qualitative data permit researchers to induce insights, which can reveal new constructs and potentially new relationships among constructs. By iterating abductively between qualitative Big Data and existing theory, researchers are given windows into theorizing supply chains that would have otherwise been closed with strictly quantitative data, such as the emergence, functioning, and adaptation of unexpected interorganizational structures and behaviors.

One of the strengths as well as challenges of qualitative Big Data is that the data are incredibly rich of contextual details, which makes them difficult to reduce and analyze. Qualitative data include not only text, but also videos, photographs, or sounds, which can be scraped from Web sites, social media, news articles, and open-source contracts. Machine learning algorithms, especially those related to topic modeling, can process these data to help researchers see patterns that would otherwise have been impossible because of the sheer volume of irreducible data.

In this paper, we first describe what we mean by qualitative Big Data. Then, we describe topic modeling, which is particularly suited for extracting insights from qualitative Big Data. This section is followed by illustrations of the application of topic modeling in two different studies in organizational theory journals, with a description of how their analysis can be

All authors are contributed equally to the development of this research paper.

[1] We should note that although the word "data" is plural, we treat the expression as a proper noun "Big Data," which is capitalized and singular.

extended to theorizing supply chains. Specifically, the methods used by Almquist and Bagozzi (2019) can examine the functioning of supply chains, and the methods used by Croidieu and Kim (2018) can examine the temporal dynamics of supply chain emergence and adaptation. We conclude the paper by outlining few cautionary notes for theorizing supply chains from qualitative Big Data.

THEORY DEVELOPMENT WITH QUALITATIVE BIG DATA

Researchers have highlighted three different qualities of Big Data: volume, velocity, and variety (3Vs) (Brinch, 2018; Constantiou & Kallinikos, 2015; George et al., 2016; McAfee et al., 2012). Volume reflects the size of the dataset, including the large number of variables and observations; velocity refers to the speed at which the data are collected; and variety represents the range of data sources including text, video, audio, images, networks, and graphics. Only volume is a necessary condition for Big Data, but it is not a sufficient condition because supply chain scholars have long generated immense datasets.

Whereas these qualities have been useful in helping to conceptualize Big Data, we believe that Big Data must also "require you to change your mind-set" (Tonidandel, King & Cortina, 2018: 525). To do so, Big Data must be sufficiently rich and contextualized to allow researchers to induce and elaborate supply chain theory that considers, for example, multiple organizations (e.g., from buyers and suppliers to logistics providers and nongovernmental organizations) with diverse connections (e.g., from material flows to contractual relationships and resource exchanges) over time.

Technology is opening up numerous sources of *qualitative* Big Data, especially through social media (e.g., email, Facebook, Twitter, Reddit), private open sharing (e.g., ResourceContracts.org, Government Open-Data, Open-Source Movements), and from scraping Web sites. Digitalization can generate qualitative data in volume and richness never before imagined. Anything that anyone has ever written is often captured through technology, leading to an explosion of new voluminous data that can reveal previously hidden insights (Boyd & Crawford, 2012; Tonidandel et al., 2018).

Supply chain phenomena can be described through words, and these descriptions can provide deeper insights into cross-level effects (e.g., between organizations, supply chains, and socio-economic or ecological environments) and trends over time—attributes of supply chains that have been difficult to empirically model in the past. In fact, text-based data can often capture salient aspects of supply chains more easily than noisy quantitative Big Data. By integrating qualitative Big Data into supply chain analysis, researchers can respond to recent calls to develop a more robust theory of the supply chain (Carter et al., 2015a, b).

There exist many open sources of qualitative Big Data for supply chain researchers, some of which we list in a dedicated Appendix. In spite of the rich data that are becoming increasingly available, few studies in supply chain management have integrated qualitative Big Data into a robust research design to build supply chain theory (notable exceptions being Ancarani et al., 2019; Chae, 2015). In the next section, we describe a research method—topic modeling—that is particularly suitable for analyzing patterns in qualitative Big Data. It can help supply chain researchers identify new constructs.

A Description of Topic Modeling

Topic modeling is a relatively new technique for analyzing qualitative Big Data, which is being used increasingly more often in organizational theory (Hannigan et al., 2019). Topic modeling can be combined with other techniques, such as network analysis, to theorize from voluminous, rich textual data. Before the advent of *unsupervised* machine learning techniques like topic modeling, Rabinovich and Cheon (2011) suggested that the effectiveness of textual analysis to develop theory depended on how well the data could match the operationalization of different constructs required for supply chain research. Organizational theorists now apply unsupervised machine learning to analyze qualitative Big Data to understand the evolution of different topics over time and across space. *Unsupervised* machine learning is completely inductive, in that the researcher knows little about the constructs and their relationships, which contrasts to *supervised* machine learning like "Nvivo automatic coding" in which the researcher must have sufficient insight to guide the coding process.

Qualitative Big Data has been argued to potentially undermine research rigor by requiring supply chain researchers to "compromise the range of their measurements, pare down the scale and scope of their models, or even limit the breadth of their review of the literature to ground their hypotheses and constructs" (Rabinovich & Cheon, 2011). However, we argue that topic modeling affords new research possibilities by revealing hidden topics. Topic modeling allows qualitative data to be synthesized in such a way to allow for interpretation, thereby allowing researchers to challenge their a priori biases. The output of topic modeling still requires researchers to induce theoretical constructs, but it does not require researchers bring significant assumptions of the data or their structures to the research process.

In the next few paragraphs, we offer a precis of the process of topic modeling. As this method is growing in interest, there are an increasing number of sources of detailed insights (e.g., DiMaggio, Nag, & Blei, 2013; Hannigan et al, 2019; Tirunillai & Tellis, 2014).

Topic modeling assumes that documents, whether they are books, reports, or articles, hold a hidden distribution of latent topics (Griffiths & Steyvers, 2004). A "*topic*" is a group of words that co-occur frequently and can be framed as theme or motif (DiMaggio et al., 2013). A document comprises a combination of multiple overlapping topics, and each topic is comprised of a list of words. Topic modeling, then, is a statistical method that renders thematic information within documents, relationships among topics and documents, and changing topical patterns in qualitative textual data (Blei, 2012; Blei, Ng, & Jordan, 2003). We describe the steps of this rendering process below.

Render Corpus or Corpora. Researchers start by first identifying a specific research question and preprocess the necessary qualitative Big Data. The researcher must identify textual data sources that meet the research parameters, such as authorship, levels of analysis, and the appropriate time frame. After the relevant textual data are sampled, scraped, and stored, the researcher decides on the textual units of analysis (called documents), which can be as small as a sentence or as long as a book chapter, that can be compared. The documents are assembled into a corpus or corpora. Documents from a single source (e.g., firms) represent a corpus, whereas documents from multiple sources (e.g., firms, industry associations, governments, and NGOs) constitute a corpora. The documents are then trimmed (nonsignificant words removed) and stemmed (different words with the same semantic content) to reduce the noise in detecting co-occurring words.

Render Topics. The researcher then applies a topic modeling algorithm on the corpus or corpora to extract co-occurring words that form topics. The researcher must decide on the optimal number of topics that the algorithm should search for to reveal the hidden distribution of constructs. The number of topics must be sufficiently broad to discover hidden constructs (validity) but also sufficiently narrow to be relevant to the researcher to answer specific research questions (accuracy). If the topics are too narrow or too broad, they may not be sufficiently distinctive semantically to help researchers theorize (DiMaggio et al, 2013; Kaplan & Vakili, 2015).

Render Theoretical Artifacts. After the researcher identifies the appropriate topics, s/he must draw out the stories behind the rendered topics that illuminate theory by considering the relationships among the topics and their context. Often times, researchers are advised to complete this final task in teams, working iteratively first alone and then in groups. Similar to traditional inductive or abductive theorizing, when inconsistencies between researchers' interpretations arise, it is important to collectively recontextualize data and discuss each other's views until discrepancies are resolved.

There are numerous benefits to topic modeling in theorizing for supply chains. It is effective in identifying topics that might be difficult to spot without unsupervised machine learning. It also helps overcome at least some of the systematic biases of researchers that may creep in with traditional qualitative research techniques that rely heavily on the interpretation of the text by the researcher, such as coding with a popular qualitative approach like the Gioia method (Gioia, Corley, & Hamilton, 2013). When done well, topic modeling involves a team of diverse researchers who can bring disparate knowledge to the table, which help to overcome personal or disciplinary biases. Finally, topic modeling requires considerable interpretation of the co-occurring words, which can stimulate new insights among researchers—insights that can be missed by simply reading the text. The onus, however, still lies on the researcher to make meaning of the topics. Sets of co-occurring words can be simply noise, or they may signal a new construct or relationship among constructs.

EXAMPLES OF THEORIZING FROM ORGANIZATIONAL THEORY USING QUALITATIVE BIG DATA AND TOPIC MODELING

Qualitative Big Data and topic modeling are only just emerging in organization theory journals and have not yet spread to supply chain journals. We describe two studies from organizational theory below that could inform supply chain research.

Examining Functioning: Almquist and Bagozzi (2019)

Description of Almquist and Bagozzi (2019). The authors sought to understand the network structure and functioning of radical environmental organizations, such as *Earth First!*, and used qualitative Big Data and topic modeling to identify the tactics these activists used to mobilize social change. Almquist and Bagozzi (2019) provided a window into the type of analysis that could help supply chain researchers understand the connections among multiple supply chain members and why and how some "supply chains persist and others expire" (Carter et al., 2015b: 95).

First, Almquist and Bagozzi (2019) drew their data from a single UK radical environmental publication, called *Do or Die*, from 1992 to 2003. A total of 10 issues were published in that period, with the first issue having only 20 pages and the last having 343 pages. The authors first decomposed each *Do or Die* issue into 12-sentence segments, called "documents," which they argued and showed is the appropriate length for modeling topics.

Second, using a relatively simple algorithm, the authors identified co-occurring pairs of environmental organizations within each document, from which they were able to develop a network map of 143 organizations and basic network statistics such as density and degree distributions. This analysis revealed the centrality of two groups of organizations: *Reclaim the Streets* and *Class War*. The authors also used structural equivalence measures and dendograms to isolate three dominant clusters within the network.

Third, the authors applied topic modeling to understand the social movement tactics and coordination mechanisms manifesting within co-occurring binary pairs of environmental organizations and within their clusters. An illustration of the final output rendered by this analysis is provided in Figure 1. The authors found fifteen topics in total, of which four were particularly salient to social movement tactics: violent protest; direct action/ecotage; occupation/camps; and international terror. They could also assign these topics to the pairs of organizations and to their clusters. For example, they found that *Oxford Earth First!* and *Class War* used violent protest more than direct action, whereas *Reclaim the Streets* and *London Green-Peace* used less violent protest, but more direct action. And, at the cluster level, they found that both clusters 1 and 2 used direct action, but cluster 1 used more violent protest.

This creative application of topic modeling allowed Almquist and Bagozzi (2019) to systematically analyze a large corpus of qualitative data to detect how diverse social movement tactics were used by different organizations in different clusters and how diverse organizations were coordinating their actions. This type of systematic analysis has been previously out of reach for qualitative researchers and helped the authors theorize reasons and patterns of activist cooperation that potentially explained the downfall of radical environmental protests in the UK in the late 1990s.

How Supply Chain Researchers can Apply Almquist and Bagozzi (2019). Taking inspiration from the work of Almquist and Bagozzi (2019), supply chain researchers could first collect lists of buyer–supplier connections from private databases such as Mergent, Compustat, Bloomberg SPLC, FactSet, or Panjiva and then augment such quantitative data with qualitative Big Data, by scraping text from reports, Web sites, news articles, and social networks like Twitter. This way, multiple types of ties—from contracts to joint projects and advocacy efforts—between diverse supply chain members—from businesses to nongovernmental organizations—could be illuminated to better describe and explain the (in)capable functioning of a supply chain. This approach to data collection and analysis based on topic modeling will allow for more complete network maps, while also recognizing diverse tactics and coordination mechanisms adopted by each pair of organizations or clusters in the large network. We believe that future studies leveraging this method will be able to describe and conceptualize aggregate supply chain behavior, with potentially surprising and revealing results on the interplay of action and structure—that is, how structure forms from action and how action flows from structure (Sydow & Windeler, 1998)—and on how complex patterns of organization-level behaviors and connections ultimately affect supply chain-level outcomes.

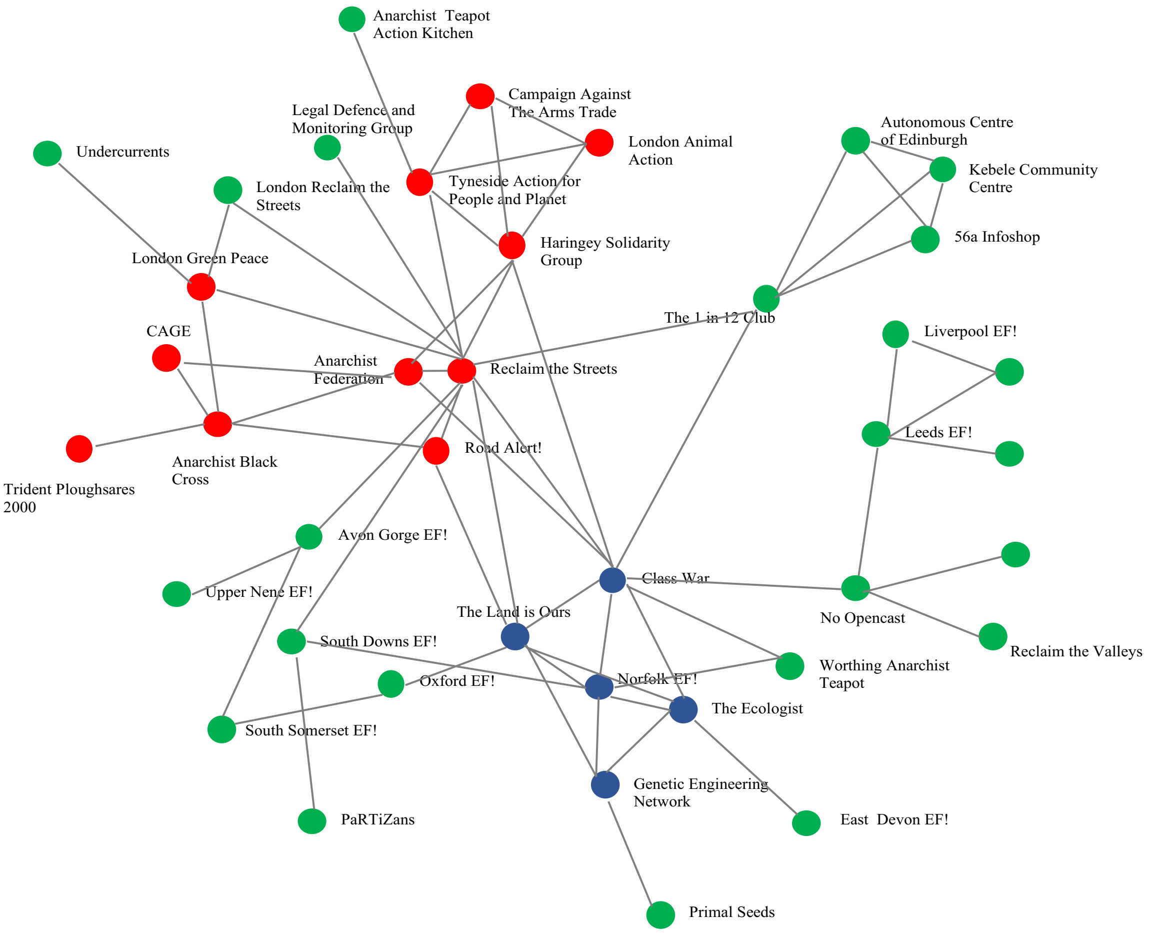
Examining Emergence: Croidieu and Kim (2018)

Description of Croidieu and Kim (2018). The authors were interested in understanding how people outside of a professional field came to be recognized as holding expert knowledge by the professionals within the field. These authors pursued this interest by analyzing the evolution of the U.S. wireless-radio-broadcasting industry, in which radio operators, or "hams" as they were called, went from being viewed as a nuisance to their knowledge being viewed over time as important to the profession and collaborating with professionals in the field, including top scientists and Nobel Prize winners. Croidieu and Kim (2018) used topic modeling, but unlike Almquist and Bagozzi (2019), focused on the temporal, evolutionary dynamics of the U.S. radio broadcasting industry.

The authors collected 2,168 documents (2.2 million words) from 1899 to 1927 from various sources, including the U.S. government, radio electrical engineers, radio corporations, and the *New York Times*. Such textual material was assumed to reflect the evolving perceptions of key actors in the early development of U.S. radio broadcasting. Unlike the previously described study, Croidieu and Kim (2018) analyzed the text *over time* to detect trends in the identified topics in relation to the evolving institutional context of the U.S. radio broadcasting industry at the beginning of the nineteenth century. They aimed to detect topical trends in order to describe the underlying mechanisms that led to the legitimizing of amateur radio operators as industry experts.

The authors went through a number of steps to detect and make sense of evolving topics in the industry, much like the application of the Gioia method in qualitative data analysis (Gioia et al., 2013). First, they coded individually and collectively the rendered topics into first-order concepts and

FIGURE 1
The Functioning of a Network of Radical Environmental Organizations—Adapted from Almquist and Bagozzi (2019). [Color figure can be viewed at wileyonlinelibrary.com]



🔴 Cluster 1	🔵 Cluster 2	🟢 Cluster 3
• Leftist environmental groups. • Associated with tactics of violent protest and direct action/ecotage. • Coordinating street protests and direct action campaigns. • High centrality in the overall network.	• Environmentalist in nature. • Associated with direct action/ecotage but less related to violent protest. • Reporting and publicizing direct action strategies, instead of active participation. • Disseminating information.	• A number of regional EF! Groups and organizing centers. • Concerned with the occupation/camps topic. • Coordinating more targeted direct action. • Offering venues where direct action activities are organized (i.e. anarchist bookstores, community centers).

second-order themes; then, they iterated multiple times between the original stories and their emerging interpretations of surfaced topics to generate new theoretical mechanisms (See Table 1 for a simplified illustration of topics and aggregate dimensions).

TABLE 1

From Topics to Theory

Automatically rendered topics and their vector of co-occurring words	First-order concepts	Second-order themes	Aggregate dimensions
Topic #391: hole screw brass drill fasten place strip thread solder requir make center machin wood mount support knob point give hold	Building a basic radio set	Forming a rudimentary competence	Mechanism 1: Build an advanced collective competence
Topic #96: book receiv includ practic work given radio instruct construct chapter principl treat modern time describ subject matter volum show design	Sharing instructional principles	Learning and sharing knowledge collectively	
Topic #47: station work amateur heard time citi mile relay ohio state coast denver distanc west record angel handl texa good district	Conducting radio relays	Developing a distinctive collective competence	

Through this iterative process, and by showing the yearly distribution of the topics across diverse stakeholder groups over 28 years, the authors were able to identify both explicit and implicit conditions of lay-expertise legitimization mechanisms that occurred conjointly in different areas of the U.S radio broadcasting industry. For example, they found that from 1899 to 1918 amateur radio operators started to build advanced competence in radio operations, which complemented the knowledge of the more academic “professionals.” From 1919 to 1927, radio operators deepened this knowledge and operated in a more public space, where they were being noticed by the general public and the professionals within the industry.

How Supply Chain Researchers can Apply Croidieu and Kim (2018). The highly contextualized, abductive process of elaborating theory has been a perennial challenge for supply chain researchers, because there are so little granular data at the requisite levels of analysis. Qualitative Big Data and topic modeling help to overcome this challenge. By exploring and visualizing text-based data, supply chain researchers can isolate anomalies that can spark the journey to elaborating new theory. Such anomalies are often difficult to detect in conventional qualitative studies because of the effort needed to collect rich primary data across multiple settings. As a result, researchers may think they might be seeing an anomaly, but they did not have the data to see if the incident appeared elsewhere or could not explore sufficiently to understand the anomaly. Qualitative Big Data, by contrast, has grain, extent, and contextual richness. It contains a large range of data, but through which the context can be somewhat preserved and meanings interpreted, allowing researchers to investigate the constructs that could influence the phenomenon of interest. Researchers can dynamically explore and reduce rich text-based data with the express purpose of exposing anomalies and with sufficient data to allow for the needed theoretical elaboration.

For example, among other relevant topics, supply chain researchers have considered how firms can sense, prevent, and respond to suppliers’ environmental and social violations (Pagell & Wu, 2009). This research stream is primarily concerned with categorizing diverse sustainable sourcing practices—such as arm’s length supplier auditing or collaborative supplier development—and explain how firms choose between them or combine them to develop more sustainable supply chains (Hajmohammad & Vachon, 2016; Villena & Gioia, 2018).

Theorizing from qualitative Big Data can help to render new categories of sustainable practices (Székely & vom Brocke, 2017) but also develop new theory on the temporal dynamics of sustainable practice emergence and legitimization. For example, taking inspiration from the work of Croidieu and Kim (2018), supply chain researchers could scrape data for all the major stakeholder groups active in electronics or apparel industries in the past 15-20 years, from global manufacturers and their key suppliers to international nongovernmental organizations and activist groups. These corpora of textual data would allow to ask why and how specific sustainable sourcing practices come to dominate the industry and be endorsed by powerful industry associations such as the Responsible Business Alliance or the International Apparel Federation, whereas others are quickly abandoned. This study would greatly contribute to our limited understanding of the temporal patterns of sustainable practice emergence and legitimization. Importantly, it could also advance supply chain practice by explaining why arm's length supplier auditing continues to be largely adopted in spite of the overwhelming evidence of its ineffectiveness (Locke et al., 2007; Porteous, Rammohan & Lee, 2015).

ADVANCING SUPPLY CHAIN THEORY THROUGH THE ANALYSIS OF QUALITATIVE BIG DATA

We believe that qualitative Big Data would be particularly helpful in theorizing supply chains as complex adaptive systems (CAS), which have been largely understudied (Nair & Reed-Tsochas, 2019). *Quantitative* Big Data has helped to study supply chains as CAS through dynamic modeling. Park et al. (2018), for example, used Compustat data to examine the impact of prior structural configurations and industry growth on the trajectory of supply chain connections, thus revealing their hidden anatomy and endogenous evolution over time. Studying supply chains as CAS via dynamic modeling, however, represents a daunting task because quantitative data are vast, granular, and decontextualized and models must be built a priori. Differently, *qualitative* Big Data and topic modeling can "speak to" such issues as supply chain emergence, functioning, and adaptation by allowing topics to emerge from large volumes of contextually rich data. To make these arguments, we describe what we mean by CAS and then explore a tentative avenue for future research.

We apply Pathak, Day, Nair, Sawaya & Kristal's definition of a CAS: a "system of interconnected autonomous firms that make choices to survive and, as a collective, the system evolves and self-organizes over time" (Pathak, Day, Nair, Sawaya & Kristal, 2007: 562). A supply chain "*system*" is simply a set of interconnected organizations, including buyers, suppliers, and support actors like logistics providers, financial institutions, and nongovernmental organizations that offer additional resources and capabilities (Carter et al., 2015b; Gualandris et al., 2015).

The "*complexity*" of the supply chain system is determined by the number, quality, and patterns of interconnections among the organizations (Choi, Dooley & Rungtusanatham, 2001; Nair & Reed-Tsochas, 2019). The bases for these connections are many, some of which are visible such as the flow of products, services, and money, and some of which are not-so-visible, such as the exchange of information and knowledge, the existence of dense social ties, and the awareness of others' competitive actions and connections (Johnson et al., 2018; Lu & Shang, 2017).

A complex supply chain system is "*adaptive*" when its internal interacting organizations co-evolve so that it most efficiently achieves its objective vis-à-vis its surrounding socio-economic or ecological environments (Kauffman et al., 2018). Over time, a system will often reach a stable state called homeostasis, so its organizations and connections continue to adjust to each other, but the magnitude of structural changes is relatively small as the system becomes locally fit. When the surrounding environment changes, elements of the system must readjust and "adapt" to the environmental changes.

Understanding supply chain emergence, functioning, and adaptation requires researchers to examine its nested, hierarchical structures: A focal firm's plants form a manufacturing network (Ferdows, 2008), which are connected to buyers and suppliers to form strategic supply chains (Ireland & Webb, 2007) that are in turn nested into a multitiered supply network (Sharma et al., 2019). Cross-level effects are important to recognize, as the behaviors and outcomes at one level of analysis (plant) may shape and be shaped by another (supply network).

Shifting the research lens from supply chains as static networks of organizations to a complex adaptive system is both theoretically and empirically daunting. Most supply chain researchers see supply chains as well-defined bounded structures and focus on a single level of analysis at a single point in time and on specific organizations, products, or services (Miller et al., 2018). For the most part, supply chain researchers have modeled average, linear, causal behavior of "small causes lead to small effects" (Nair & Reed-Tsochas, 2019: 89). Whereas a bilateral exchange between two organizations is relatively easy to see and measure, it is much more challenging to collect and analyze data that reveal patterns among interconnected elements across levels and over time. It is for these reasons that Carter et al (2015b) and Kauffman et al (2018) mention that very little research has empirically examined supply chains as complex, living systems. A shift to CAS requires supply chain researchers to deliberately choose the organizations, level(s) of analysis, and connections over time, which imposes a heavy data burden. We believe that qualitative Big Data can help meet this need.

For example, scholars could produce insights as to adaptability of supply chain systems by exploring emerging practices and structures in a circular bioeconomy (D'Amato et al., 2017). Such a study could reveal the generative mechanisms responsible for developing novel supply chain connections among players within and across diverse industries. For example, scholars could collect articles and reports from large and small players active in an agrifood supply chain system in Europe over time. Key organizations may include, but are not limited to, farmers that upcycle crop waste into bioplastic (Checchini, 2017), food processors that try to limit food loss in their operations while also sourcing the waste streams of other organizations (Bansal & McKnight, 2009), large retailers that recirculate food waste and food packaging waste for alternative uses in other industries, non-governmental organizations that act as circularity brokers (Ciulli, Kolk & Boe-Lillegraven, 2019), and governments that stimulate innovation and change the "rules of the game."

This study could reveal new co-evolution mechanisms between supply chain members that are spatially and industrially distant, yet connected through the transfer of goods in the emerging circular bioeconomy. It could also help delve deeper into the adaptation process of the supply chain to changes in the broader socio-economic environment by specifying how diverse members manage uncertainties and respond to unanticipated events, such as the adoption of the 2015 EU action plan for the circular economy. Kauffman et al (2018) describe the conditions under which buyers and suppliers can "tinker" their interorganizational connections, so they are more adaptive to unanticipated events in the environment, but these authors do not offer insights into the temporal processes by which this happens. The analysis that we propose, which is inspired by the works of Almquist and Bagozzi (2019) and Croidieu and Kim (2018), could reveal the specific mechanisms that make complex supply chain systems more (or less) adaptive.

Supply chain researchers have the opportunity to glean new theoretical insights from qualitative Big Data, slicing and dicing in different ways, similar to topic modeling analysis previously illustrated. However, the high volume and variety of data for a limited set of observational units sometimes mean that the researcher must not only have the generative and interpretative skills of an inductive researcher but also the deductive technical skills to manage datasets that have both grain and extent.

CAUTIONARY NOTES FOR SUPPLY CHAIN RESEARCHERS THEORIZING FROM QUALITATIVE BIG DATA

Although qualitative Big Data offers new research opportunities, it also has limitations that can become research traps. We outline several below.

Ethical Issues

Despite the potential of qualitative Big Data to contribute to supply chain theory, ethical issues and data privacy pose serious concerns (Boyd & Crawford, 2012; Tonidandel et al., 2018). Big Data is often generated by the continuous accumulation and storage of information that does not have a specific purpose or a purpose different from the research endeavor. When the data are acquired and analyzed, researchers are often not sure what they will find and study. Or, they may be able to access data that were not intended for public use, breaching ethical or legal concerns (Simsek et al., 2019). For example, Internet-enabled devices allow companies to "see" and "hear" into peoples' homes, learning about what is consumed, by whom and how. Just because a corporation has planted a device in peoples' homes does not mean that the data should be analyzed, especially by third parties.

Debates on the practices and the standards regarding demands of protection of individual data rights are ongoing (Simsek et al., 2019). It is difficult to define boundaries of "public" data, the people who have the rights to consume them, and the types of analysis that are deemed appropriate. Until these standards are set, corporations are assuming that the data that they collect through their activities are theirs to use, analyze, and distribute.

To address ethical and privacy issues, researchers must rely on their own internal ethical compass to determine how the data should be used. Researchers could consult with their university's ethics boards to ensure that individual well-being, organizational well-being, and societal well-being are protected if they are collecting or analyzing primary individual-level data.

Valid Constructs

Researchers should always validate the topics derived from qualitative Big Data to exclude potential veracity issues—that is, "garbage in and garbage out" (Demchenko et al., 2013)—and confirm that topic modeling has generated topics that yield meaningful constructs. Co-occurring words can be gibberish, and it is important for researchers to impose meaning to those words that in some way speak to and extend existing theory.

For instance, in a study that examined the value of financial analyst information discovery and interpretation roles for investors, Huang et al. (2018) ran three validation tests for the topic modeling outputs. First, the authors manually reviewed the 20 most frequent words in each topic, read relevant sentences in the original text, and put a short label for each topic. By doing so, they confirmed that topic modeling could distinguish the underlying financial content of the topics. The second validation test compared the temporal variation in the amount of discussion devoted to key topics across documents (i.e., analysts' reports) in coincidence with major industry and economy-wide shocks. This test ensured that variation in the rendered topics over time would be unsurprisingly related to key exogenous shocks; for example, discussion concerning topics of "growth" and "mortgage origination" decreased in coincidence of the financial crisis in 2008 while discussion on topics like "real estate loans" and "deteriorating performance" increased substantially over the same timeframe. Lastly, for a small subsample of documents, the authors compared the topic assignment to individual sentences by topic modeling with manual topic assignment by a human coder, which showed satisfactory levels of consistency (60%-69% inter-rater reliability scores).

Spurious Relationships

Qualitative Big Data requires a large volume of data to be compressed for interpretability, which can strip away rich context. By losing the context, researchers may see relationships among constructs that are meaningless (Bansal, Kim & Wood, 2018; Grimmer & Stewart, 2013). Researchers need to discriminate between spurious, bogus relationships yielded by topic modeling by returning to theory and the actual phenomena to see if the insights being rendered from qualitative Big Data are explainable. Both of the articles that we illustrated rely on existing theories to elaborate their findings. Moreover, researchers can also consider triangulating small-scale qualitative data from field interviews and surveys with the insights from qualitative Big Data (Ford et al., 2016). Discussions over rendered topics with industry experts will lead to more reliable theoretical artifacts. While applying topic modeling on qualitative Big Data will uncover clandestine information, insights from small data will help to reliably interpret insights from Big Data.

CONCLUSION

The study of supply chains has been hampered by the need for large volumes of rich, granular data that include diverse organizations, cover diverse types of connections, and show emergence and functioning over time. Qualitative Big Data provides an opportunity to theorize supply chains more fully through topic modeling, especially to understand supply chains as complex adaptive systems. Yet, qualitative Big Data has been largely overlooked by supply chain researchers. We hope this paper offers supply chain researchers inspiration for further research into supply chains. Qualitative Big Data has the opportunity for supply chain researchers to shine a spotlight on a landscape that has been previously hidden from view. We are excited about the prospects.

REFERENCES

Almquist, Z. W., & Bagozzi, B. E. (2019). Using radical environmentalist texts to uncover network structure and network features. *Sociological Methods & Research*, 48, 905–960.

Ancarani, A., Di Mauro, C., Legenvre, H., & Cardella, M. (2019). Internet of things adoption: A typology of projects. *International Journal of Operations and Production Management*. https://doi.org/10.1108/IJOPM-01-2019-0095 [Epub ahead of print].

Bansal, P., Kim, A., & Wood, M. O. (2018). Hidden in plain sight: The importance of scale in organizations' attention to issues. *Academy of Management Review*, 43, 217–241.

Bansal, P., & McKnight, B. (2009). Looking forward, pushing back and peering sideways: Analyzing the sustainability of industrial symbiosis. *Journal of Supply Chain Management*, 45, 26–37.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55, 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15, 662–679.

Brinch, M. (2018). Understanding the value of big data in supply chain management and its business processes: Towards a conceptual framework. *International Journal of Operations & Production Management*, 38, 1589–1614.

Carter, C. R., Meschnig, G., & Kaufmann, L. (2015a). Moving to the next level: Why our discipline needs more multilevel theorization. *Journal of Supply Chain Management*, 51, 94–102.

Carter, C. R., Rogers, D. S., & Choi, T. Y. (2015b). Toward the theory of the supply chain. *Journal of Supply Chain Management*, 51, 89–97.

Chae, B. K. (2015). Insights from hashtag# supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research. *International Journal of Production Economics*, 165, 247–259.

Choi, T. Y., Dooley, K. J., & Rungtusanatham, M. (2001). Supply networks and complex adaptive systems: Control versus emergence. *Journal of Operations Management*, 19, 351–366.

Ciulli, F., Kolk, A., & Boe-Lillegraven, S. (2019). Circularity brokers: Digital platform organizations and waste recovery in food supply chains. *Journal of Business Ethics*, 00, 1–33.

Constantiou, I. D., & Kallinikos, J. (2015). New games, new rules: Big data and the changing context of strategy. *Journal of Information Technology*, 30, 44–57.

Croidieu, G., & Kim, P. H. (2018). Labor of love: Amateurs and lay-expertise legitimation in the early US radio field. *Administrative Science Quarterly*, 63, 1–42.

D'Amato, D., Droste, N., Allen, B., Kettunen, M., Lähtinen, K., Korhonen, J., . . . & Toppinen, A. (2017). Green, circular, bio economy: A comparative analysis of sustainability avenues. *Journal of Cleaner Production*, 168, 716–734.

Demchenko, Y., Grosso, P., De Laat, C., & Membrey, P. (2013). Addressing big data issues in scientific data infrastructure. In *2013 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 48–55). San Diego: IEEE.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics*, 41, 570–606.

Ferdows, K. (2008). Managing evolving global production networks. In R. Galvan (Ed.), *Strategy innovation and change: Challenges for management* (pp. 149–162). Oxford, UK: Oxford University Press.

Ford, J. D., Tilleard, S. E., Berrang-Ford, L., Araos, M., Biesbroek, R., Lesnikowski, A. C., … & Bizikova, L. (2016). Opinion: Big data has big potential for applications to climate change adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 10729–10732.

George, G., Osinga, E. C., Lavie, D., & Scott, B. A. (2016). Big data and data science methods for management research. *Academy of Management Journal*, 59, 1493–1507.

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. *Organizational Research Methods*, 16, 15–31.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(suppl 1), 5228–5235.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21, 267–297.

Gualandris, J., Klassen, R. D., Vachon, S., & Kalchschmidt, M. (2015). Sustainable evaluation and verification in supply chains: Aligning and leveraging accountability to stakeholders. *Journal of Operations Management*, 38, 1–13.

Hajmohammad, S., & Vachon, S. (2016). Mitigation, avoidance, or acceptance? Managing supplier sustainability risk. *Journal of Supply Chain Management*, 52, 48–65.

Hannigan, T. R., Haans, R. F. J., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., … Jennings, P. D. (2019). Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals*, 13, 586–632.

Huang, A. H., Lehavy, R., Zang, A. Y., & Zheng, R. (2018). Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science*, 64, 2833–2855.

Ireland, R. D., & Webb, J. W. (2007). A multi-theoretic perspective on trust and power in strategic supply chains. *Journal of Operations Management*, 25, 482–497.

Johnson, J. L., Dooley, K. J., Hyatt, D. G., & Hutson, A. M. (2018). EMERGING DISCOURSE INCUBATOR: Cross-sector relations in global supply chains: A social capital perspective. *Journal of Supply Chain Management*, 54, 21–33.

Kaplan, S., & Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal*, 36, 1435–1457.

Kauffman, S., Pathak, S. D., Sen, P. K., & Choi, T. Y. (2018). Jury rigging and supply network design: Evolutionary "Tinkering" in the presence of unknown-unknowns. *Journal of Supply Chain Management*, 54, 51–63.

Locke, R. M., Qin, F., & Brause, A. (2007). Does monitoring improve labor standards? Lessons from Nike. *ILR Review*, 61, 3–31.

Lu, G., & Shang, G. (2017). Impact of supply base structural complexity on financial performance: Roles of visible and not-so-visible characteristics. *Journal of Operations Management*, 53, 23–44.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. *Harvard Business Review*, 90, 60–68.

Miller, J. W., Ganster, D. C., & Griffis, S. E. (2018). Leveraging big data to develop supply chain management theory: The case of panel data. *Journal of Business Logistics*, 39, 182–202.

Mishra, D., Gunasekaran, A., Papadopoulos, T., & Childe, S. J. (2018). Big Data and supply chain management: A review and bibliometric analysis. *Annals of Operations Research*, 270, 313–336.

Nair, A., & Reed-Tsochas, F. (2019). Revisiting the complex adaptive systems paradigm: Leading perspectives for researching operations and supply chain management issues. *Journal of Operations Management*, 65, 80–92.

Pagell, M., & Wu, Z. (2009). Building a more complete theory of sustainable supply chain management using case studies of 10 exemplars. *Journal of Supply Chain Management*, 45, 37–56.

Park, H., Bellamy, M. A., & Basole, R. C. (2018). Structural anatomy and evolution of supply chain alliance networks: A multi-method approach. *Journal of Operations Management*, 63, 79–96.

Pathak, S. D., Day, J. M., Nair, A., Sawaya, W. J., & Kristal, M. M. (2007). Complexity and adaptivity in supply networks: Building supply network theory using a complex adaptive systems perspective. *Decision Sciences*, 38, 547–580.

Porteous, A. H., Rammohan, S. V., & Lee, H. L. (2015). Carrots or sticks? Improving social and environmental compliance at suppliers through incentives and penalties. *Production and Operations Management*, 24, 1402–1413.

Rabinovich, E., & Cheon, S. (2011). Expanding horizons and deepening understanding via the use of secondary data sources. *Journal of Business Logistics*, 32, 303–316.

Sharma, A., Pathak, S., Borah, S. B., & Adhikary, A. (2019). Is it too complex? The curious case of supply network complexity and focal firm innovation. *Journal of Operations Management*. https://doi.org/10.1002/joom.1067 [Epub ahead of print].

Simsek, Z., Vaara, E., Paruchuri, S., Nadkarni, S., & Shaw, J. D. (2019). New ways of seeing big data. *Academy of Management Journal*, 62, 971–978.

Sydow, J., & Windeler, A. (1998). Organizing and evaluating interfirm networks: A structurationist perspective on network processes and effectiveness. *Organization Science*, 9, 265–284.

Székely, N., & vom Brocke, J. (2017). What can we learn from corporate sustainability reporting? Deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique. *PLoS ONE*, 12, e0174807.

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, 51, 463–479.

Tonidandel, S., King, E. B., & Cortina, J. M. (2018). Big data methods: Leveraging modern data analytic techniques to build organizational science. *Organizational Research Methods*, 21, 525–547.

Villena, V. H., & Gioia, D. A. (2018). On the riskiness of lower-tier suppliers: Managing sustainability in supply networks. *Journal of Operations Management*, 64, 65–87.

Dr. Tima (Pratima) Bansal (DPhil from Oxford University) is a Canada Research Chair in Business Sustainability at the Ivey Business School, Western University. Dr. Bansal researches the dimensions of time, space and scale in business strategy to understand business sustainability. She applies both qualitative and quantitative methods. She has published widely in management journals, including *Academy of Management Journal, Academy of Management Review*, and *Strategic Management Journal*. She also sat as a Deputy Editor and Associate Editor of the *Academy of Management Journal* for a total of six years.

Jury Gualandris (PhD in Economics and Technology Management, Universita degli Studi di Bergamo (Italy) is an Assistant Professor in the Operations Management and Sustainability groups at Ivey Business School, Western University. Dr. Gualandris uses operational and organizational lenses and empirical methods to study the development of supply chains that strive to operate within the thresholds of socio-ecological systems. This research has been featured by 'The Conversation' and the 'Network for Business Sustainability' and has appeared in top operations management journals like the *Journal of Operations Management*, the *Journal of Supply Chain Management*, and the *International Journal of Operations and Production Management*, among many others. Jury serves in the Editorial Review Board of these journals and leads the Circular Economy Priority of the Building Sustainable Value (BSV) Centre at Ivey.

Nahyun Kim (MS in International Business and Strategy, Korea University (South Korea)) Nahyun Kim is a PhD candidate in Sustainability at the Ivey Business School, Western University. She researches corporate communications on social and environmental issues, with special emphasis on vocabularies and grammatical structures in public firms' corporate sustainability reporting. Her dissertation uses topic modeling to inductively explain how language orients firms' attention in the aftermath of financial crisis.

APPENDIX

QUALITATIVE BIG DATA SOURCES FOR SUPPLY CHAIN SCHOLARS

This appendix presents qualitative Big Data sources for supply chain scholars. While the list is not exhaustive, it provides evidence of the growing opportunities to theorize supply chains through qualitative Big Data.

First, ResourceContracts.org and OpenLandContracts.org are online repositories of publicly available contracts covering oil and gas, mining, agriculture, and forestry but also community–investor and community–government contracts. These repositories are developed in partnership with the World Bank, the Natural Resource Governance Institute (NRGI), and the Columbia Center on Sustainable Investment. Collectively, these repositories cover more than 2,500 contracts across more than 90 countries, both developed and developing, stretching over 30 years of data. ResourceContracts.org covers contracts for more than 49 extractive resources, from hydrocarbons to gold and diamonds, the majority of which are from the Democratic Republic of Congo. OpenLandContracts.org covers contracts for more than 54 commodities, from timber to palm oils and cotton. Contracts from these two repositories involve governments, businesses, and local communities, and vary from joint ventures, services, and concession agreements. Each contract features detailed textual descriptions of exchange partners' full legal names, sites location, start date and term, work and investment commitments, governing law, arbitration and dispute resolution, auditing mechanisms and restrictions related to local employment, environmental protections, and confidentiality.

Second, the Government of Canada is now fully disclosing procurement data at both federal and provincial levels, which include details about 646k contractual records over the past 15 years. Each record indicates the end user entity (e.g., the Department of National Defense, Fisheries and Oceans Canada), the reference number and the original documentation of the tender, the supplier full legal name, the supplier country and postal code, the supplier employee count, total contract value in Canadian dollars, details on the underlying good or service (e.g., material handling equipment, fishing vessels, software, and logistic services), and the type of solicitation and trade agreement adopted. Contractual data are updated monthly and organized by fiscal years. Similar to the Government of Canada, the European Union publicly discloses "Tenders Electronic Daily TED" data covering public procurement for the large European Economic Area. This data source includes over 128,000 contracts categorized by buyer, place of delivery, business sector, and type of business opportunity (e.g., expression of interest and design contests). Available data per contract contain the most important fields from the contract notice and contract award notice, such as who bought what from whom, when, at what price, and through what awarding procedure and criteria. For example, an advanced search for invitations and contracts with publication date occurring between July 1, 2019 and July 31, 2019 produced over 300 fully documented and easily downloadable contractual records.

With digitalization, we see the rise Open-Source Movements. For example, the Open Compute Project (OPC) Foundation, originally launched by Facebook, Intel, Rackspace, and Goldman Sachs, has inspired the rise of a movement in the hardware space that aims to bring about the same kind of transparency we have seen in open-source software. On the OPC Web site, there are 14 active projects, with some of them having up to 4 subprojects. Both projects and subprojects are fully documented with rich textual data and videos describing project scope, leadership, and membership over time. The OPC Web site and its hundreds of related Wiki pages publicly disclose a large volume of detailed technical reports tagged by project, leadership, and time, providing details on the evolution of the project, its governance, and outcomes, including detailed descriptions of the diverse milestones achieved over time.