Beyond the university:
Mapping higher education after data science

University of Oxford

A thesis submitted for the degree of

Master of Science

2017

Abstract

The system of higher education has experienced a series of transformations in recent decades, characterized by an acceleration of new institutional forms and arrangements as well as new information and communication technologies appearing in the field. However, despite repeated calls within the field of sociology of education to better theorize the relationship between technology, society, and education, a conceptually rigorous basis for conducting such study has not yet crystallized. At the same time, the implications of ongoing formal transformations in the system continue to be felt, especially in the arena of data science education, which displays an especially diverse—and disruptive—mix of educational practices today. This study therefore investigates transformations in higher education using data science education as a case, pursuing two sub-goals simultaneously: first, to advance the sociological understanding of data science education; and second, to innovate a conceptual/methodological framework which enables a new approach in sociology of education for the study of technology, society, and education. Each goal addresses current gaps in respective literatures, even as each also enables the other to be performed. A multiple case study approach is adopted to do so, using an integrated conceptual foundation built on the work of Basil Bernstein and Manuel Castells, and featuring a novel, network graph data management solution developed for the study. By synthesizing and applying case findings to the macro-level environment of U.S.-based higher education, the study concludes with recommendations for that policy context and points to possible further development for study of social transformations in knowledge generally.

Acknowledgements

I thank Class Central (https://www.class-central.com/) as well as Institute for Advanced Analytics, North Carolina State University (http://analytics.ncsu.edu/) for their gracious approval and support in providing me with raw data for use in the study.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

Into the twentieth century, the university was the privileged producer and distributor in liberal democracies of science and other discourses of truth, making it crucial for understanding the relationship between knowledge and power in society. Beginning in the 1990s, however, transformations in information and communications technology, higher education, and society generally have threatened this arrangement, for example by changing the delivery of education via new mediums and technologies; the production of education via increasingly complex and diverse networks of actors; and the legitimation of educational knowledge via sources besides the academy. At the same time, and even as the sociological *phenomenon* of education remains central to the (re)production of knowledge in society, only limited progress has been made in sociology of education to conceptualize and research the relationship between ICTs, education, and society in a rigorous way. In this study, I therefore propose an integrated conceptual and methodological framework to begin to address this gap, adopting a multiple case study research design in order to do so using the case of U.S.-based data science education.

Data science education is an excellent case for this task for several reasons. Interest in the field of data science has grown steadily since at least 2012, when Harvard Business Review famously declared “data scientist” to be the “sexiest job of the 21st century" (“data science - Google Trends,” 2017; Davenport & Patil, 2012). Since that time, data science has been linked to the economic prosperity of nations, companies, and individuals, both in terms of driving social change through data science practice, and in terms of creating new categories of jobs and companies. As a field, data science has in turn relied fundamentally on investment, innovation, and growth in educational offerings designed to produce new data scientists. These educational offerings have grown in volume and diversity, especially in the United States, both in terms of the forms that educational activities take—from “bootcamps,” to Massively Open Online Courses (MOOCs), to Master’s degrees—and in terms of the types of institutions involved in producing education, from universities, to technology companies, to informal collaboration networks. There is also a gap in literature published about data science education to date, where attention has been paid to topics such as curriculum design and instructional technique, while no study has evaluated either the full diversity of *forms* of activities and institutional arrangements which characterize the field, and no study has evaluated the *sociological meaning* of data science education as a site of transformation in higher education and knowledge.

This combination of high social impact and theoretical significance—marked by rapid, widespread pedagogic and institutional disruption—together makes U.S.-based data science education a socially important topic of study, even as it also makes it an excellent case for assessing the changing role and practices of higher education in society and the economy generally. The goal of this study is then to take these two aims together, each to enhance the other: to address the need for a conceptual and methodological framework to help study shifts in higher education generally, and—*as well as in order to*—investigate emerging practices in U.S.-based data science education. Reframed as problems around which to organize theory, methodology, and analysis, these aims can be restated as follows:

1. In the case of data science education, what patterns of institutional arrangement are emerging and what do these patterns mean? (How, indeed, are we to determine what they mean?)
2. Sociologically, what’s happening in the field of higher education overall, given the patterns and conceptual orientation identified in the study of data science education?

Here, I propose that a conceptual framework for studying the above can be located at the intersection of the work of two theorists who, each in their own way but especially in complement, are especially suited to the complex, multi-level field. Namely, I propose that work in the sociology of education by Basil Bernstein, combined with the social theory of Manuel Castells, is uniquely helpful in closing the conceptual/methodological gaps identified above.

To implement this framework, the research design adopts a multiple case study format divided into two levels of analysis. First, I conduct a meso-level network analysis of data science education *actors* and *activities*, paying special attention to the production and distribution of education via new network arrangements. Next, I conduct a micro-level case analysis of five individual educational "contexts" (a specialized term in Bernstein's vocabulary) chosen for their representation of a diverse combination of actor and activity types as well as network patterns. Finally, I synthesize cross-level insights into a typology of generalized pedagogic "discourses", applying this to the macro-level environment of higher education in the United States. In support of this process—and, again, as a distinct goal in its own right—I develop an integrated research approach that translates the Bernsteinian/Castellsian conceptual framework into a methodological framework, utilizing a bespoke graph database management system as well as a novel conceptual instrument developed for the study.

In summary, this study is exploratory, but concerned with concept formation; qualitative, but also computational; and flexible, but structured by a multiple case study frame, guided throughout by a dual commitment to the two "problems" stated above. The study is also guided fundamentally by Bernstein's unique approach to research—the details of which are discussed in the following chapter, after a brief review of data science education literature.

2. Literature Review

Introduction

This chapter begins by reviewing academic publications about data science education, identifying common themes and an overarching gap which points to the need for study using new disciplinary approaches. It then establishes a foundational body of theory for accomplishing this using the conceptual framework of Basil Bernstein, a British sociologist who worked in sociology of education and whose theory and methodology provide unique benefits for dealing with the challenges posed by the study. This results in the emergence of a second gap, however, in sociology of education generally and Bernstein's theory specifically, related to institutionally and technologically diverse practices in higher education. The chapter concludes by proposing an approach to overcoming this second gap by extending Bernstein's framework using the work of Manuel Castells. This conclusion forms the basis for addressing one of the study's two key goals, of developing a conceptual/methodological framework for studying the contemporary field of technology/society/education.

Data Science Education

In a query of academic literature published about "data science" and education, I uncovered one article from 2001 (Cleveland) and fifteen from 2012 or later. Over the course of the five years that cover most of these publications, the articles reflected involvement by a growing field of disciplines and institutional actors, starting in statistics (Cleveland, 2001; Finzer, 2013; Baumer, 2015; Donoho, 2015; Hand, 2015; Veaux et al., 2017), but widening to include mathematics and computer science (Cassel & Topi, 2015; Cassel, et al., 2016; Eybers & Hattingh, 2016; Song & Zhu, 2016; Culler, 2017), library and information science (Stanton, et al., 2012; Varvel, et al., 2012; Tang & Sae-Lim, 2016 & 2017), and critical data studies (Neff, et al., 2017). Involvement also grew from individual departments and societies, to institutions including the National Science Foundation (NSF) (Cassel & Topi, 2015; Neff, et al., 2017), Association of Computing Machinery (ACM) (Cassel & Topi, 2015; Cassel, et al., 2016), and private groups such as Bell Laboratories (Cleveland, 2001) and Microsoft Research (Culler, 2017). The papers also grew in scope over this period, from reviews of individual courses or program curricula, to large-scale, state-funded, collaborative studies aimed at guiding macro-level pedagogic change.

Reviewing each paper in turn, I found that none focused on the production of data science education as a sociological object of study *per se*, but rather, all approached the topic of data science education from the perspective of designing or assessing pedagogic practice, for example in terms curriculum design or instructional methods. Some common features in this regard were to proclaim a need for interdisciplinarity and practice-based teaching methods; to advise a preference for practical application, if not industry use/involvement as such; and to reflect a sense that the field is socially and economically important, with one NSF-funded study citing future data scientist labor shortages as a problem to be resolved through educational change (Cassel & Topi, 2015). Only one paper (Neff, et al., 2017) analyzed the social construction of data science as a field of practice, arguing for more nuanced and critical assessment of the field to be taken into consideration in curriculum design; but even this did not problematize the production of data science education *per se*. I also found that, out of four studies that provided a meta-analysis of data science education programs, none defined the field of data science education to be inclusive of non-academic settings, nor did any seem to use network analysis techniques in their review of programs and institutions (Stanton, et al., 2012; Song & Zhu, 2016; Tang & Sae-Lim, 2016 & 2017; Veaux et al., 2017).

Across the fifteen studies, a shared discourse seems to operate which (1) understands the academy to be the legitimate source or at least arbiter of data science education, and (2) views data science as possessing an exceptional quality which legitimates pedagogic and institutional change in the higher education system. Through this frame, the studies can be read as a set of competing efforts by various groups or segments of the academy to discursively *claim* data science, i.e. to install a discipline or set of actors as the legitimate seat of data science knowledge production. In contrast, the frame of this study could be understood as seeking to problematize the given-ness of such a discourse as is represented in these studies: to de-center the academy as the sole producer of data science education and (or in order to) investigate the power and control relations that attend to changes in institutions and pedagogy. Put another way, I claim that by studying data science education as a sociological phenomenon, rather than approaching it prescriptively, then a critical, reflexive gap can be filled in the current data science literature. Such an approach is, precisely, the aim of sociology of education, therefore to proceed further, some conceptual and methodological grounding in this literature is required. I propose that a "Bernsteinian" approach is best suited to the unique challenges and goals of this study.

Basil Bernstein and Pedagogic Discourse

Basil Bernstein was a British sociologist whose sustained project in the sociology of education lasted from the 1960s until his death in 2000. His conceptual framework comprised multiple interconnecting concepts but no single theory as such; Bernstein rather believed that "the work of theory" was "to systematically work 'to and fro' between the question at its most general level and at its most specific through conceptualizing the transformations between levels" (Moore, 2013, 97). For Bernstein, one of these "questions at a most general level" was to understand how power and control were produced and distributed in educational settings in terms of what he called “pedagogic discourse.” He defined this concept as “the principle by which other discourses are appropriated and brought into a special relationship with each other” within a given educational context (2000, 32; quoted in Moore, 2013, 162). These “other discourses” in turn comprise a body of rules and “codes” which determine, for example, the selection and pacing of curriculum, as well as more abstract concepts such as “modes” of knowledge transmission. Much of Bernstein's project was then directed at elaborating these rules and codes of pedagogic discourse, as well as situating and connecting the discourse overall within the broader social context via additional conceptual tools and layers.

Following Bernstein, to study pedagogic discourse is therefore to study the interaction of codes and rules in a given context, and to do so in part by engaging with a process of deciphering the power and control relations at play within and across contexts, if not also “behind” them. For example, Bernstein states that “ideological positioning and oppositioning are realized in, transmitted, and legitimated by classification and framing rules,” where these classification and framing rules define features such as “hierarchy,” i.e. the structure of interaction between teachers and students; “sequencing,” i.e. the content, ordering, and pacing of curriculum; and “evaluation,” i.e. the criteria for student assessment (Bernstein, 2009, 101; quoted in Moore, 2013, 79-80). In Bernstein’s framework, then, basic features of pedagogic practice are always implicated in a larger, multi-level apparatus, so that the generation of empirical research at a micro-level of individual discourse thus always links to and drives development of more abstract analysis at higher levels. At the highest level, for example, Bernstein defined a pedagogic “device” which produces pedagogic discourses, providing their “intrinsic grammar” and so regulating them “fundamentally” (Bernstein, 2000, 28, quoted in Maton, 2013, 154). I argue, then, that Bernstein’s conceptual framework, as well as his flexible approach to constructing research overall, together support the distinct research “problems” defined in the introduction of this study, as well as the gap identified above in the data science education literature. However, some work is yet required to clarify precisely how Bernstein’s framework should be operationally mapped onto the emerging contexts under investigation.

The second gap then emerges here, in two parts. First, Bernstein never defined a way of registering variously digital, networked, or “online” educational activities within his framework, though he did point to an approach to doing so. Second, the role of technology and networks in education has continued to be undertheorized generally, meaning that the work of extending Bernstein’s project as proposed here, in turn requires working at the edge of generally agreed-upon theory in sociology of education. The implication of doing this for the field generally should then be to point to a way of adapting sociology of education to the study of networked, digital education—even without necessarily needing to commit to Bernstein’s project.

With respect to the gap in Bernstein's work: research conducted to date using Bernstein’s framework have focused only on narrow applications or aspects of technology/networks in education, or else on components of the framework rather than the whole. For example, studies have investigated “hypertext” and pedagogic discourse (Tyler, 2001); teacher practices related to e-learning (Robertson, 2007 & 2008); student experiences of technology (Bennett & Maton, 2010); and the fields of “cybereducation” (Menchik, 2004) and educational technology (Czerniewicz, 2010). Bernstein's work has also been applied to institutional change, such as analysis of pedagogic reform (Sriprakash, 2011), curriculum change policy (Stavrou, 2016), teacher experiences during a school change program (Sung, et al., 2016), and even of international collaboration networks of universities (Little, Abbas, & Singh, 2016), but these have largely been applied in traditional settings for which technology/networks did not figure in quite the sense taken by this study.

With respect to the general literature: the sustained work of Selwyn (2010, 2014, 2015)—especially Selwyn & Facer (2014)—especially points to how an "educational technology" discourse has dominated more nuanced readings of the relationship between technology, society, and education for the past few decades. A sustained example of this is discourse that arose in the late 1990s about an anticipated "virtual university". Developed as universities sought to determine how to make use of the emerging World Wide Web, this discourse produced strong criticism, notably in a series of articles by Noble (1998a, 1998b, 1998c, 1999) which argued that virtual universities were fundamentally "about making money," so that faculty would inevitably be made redundant and students left with "cyber–counterfeit" educational experiences (1998a). A less pessimistic narrative developed as well, for example Robins and Webster questioned whether "nostalgia for [the] national-liberal university is misplaced, [and] even disabling" of the field's ability to imagine a possible "global-cosmopolitan university" online (2002, 17). Both narratives have persisted into the present, for example in claims about the "Uberfication of the university" (Hall, 2016) or its "neoliberal takeover" (Busch, 2017), as well as calls to build new "imaginations" that might chart a new vision for an "ecological" university (Barnett, 2013). Throughout the development of these narratives, the use of digital, networked technology in the delivery as well as production of higher education has anyway continued at a strong pace, so that despite notable efforts to theorize e.g. the "disrupting" capacity of technology for universities (Lucas, 2016), no rigorous conceptual frame has yet been offered for evaluating the role of technology in sociology of education, *per se*. This gap, registered in the present study as a lack of strong conceptual footing for mapping sociological theory onto the emerging field, has the broader implication of limiting the ability of sociology of education to offer rigorous critique or "imaginations."

In response to these combined challenges, I argue that Bernstein's conceptual framework, extended by the "network society" theory of Manuel Castells, and with additional conceptual support from institutional economics, provides a strong program for studying sociology of education and technology generally, as well as the case of U.S.-based data science, specifically. The key to this resolution is the meaning of Bernstein's concept of "contexts," which requires further explanation.

Networked Contexts

While codes, discourses, and devices might be considered the analytical object of study in Bernstein's body of work, "contexts" serve as the ground for their analysis. Often used informally to refer implicitly to a classroom environment, Bernstein defined the concept as follows:

> What counts as a context depends not on relationships within but on relationships between contexts. The latter relationships, between, create boundary markers whereby specific contexts are distinguished by their specialized meanings and realizations. (2009, 97)

This definition is both functional and relational: contexts are defined by what they *do*, which is to realize pedagogic discourse at the level of particular environments, social actors, technologies, etc., *in terms of the boundaries that the context therefore manifests* in relation to other contexts. Put another way, contexts are to some extent co-constituted in relation to each other.

The traditional university classroom context manifested these boundaries physically: a physical classroom bounded the learning environment, at the micro-level of the context; the department or college bounded the production environment, in its meso-level dimension; and the "national-liberal" education system overall bounded the ideological environment, at a macro-level. In contrast, the field of higher education today is characterized by seemingly continuous disruption at all three of these levels. Introducing an online degree program to the "traditional" context above: digital course technology blurs the boundaries of the learning environment (micro), the role that technology companies play in providing or administrating the course technology blurs the boundaries of the production environment (meso), and the institutional ecology then defined and expressed by the ascendant role of education technology companies in determining pedagogic practice then alters the ideological environment (macro). Two distinct forms of "network" are involved in effecting, or enabling, the disruptions in this example: digital networks, and "social" networks of actors collaborating to produce education.

Manuel Castells's writings on the network society synthesizes both these digital and "social" notions of network into a single theoretical framework, one that assists in mapping out Bernsteinian "contexts" and interpreting the resulting findings while yet not imposing a network theory of sociality *per se*. That is, in contrast to perspectives such as Actor-Network Theory, which claims that social actors and social reality are metaphysically co-constituted through networks; or Pentland's "social physics", in which big data combined with network science is supposed to replace the need for social theory; Castells instead claims only that digitally mediated networks constitute "the specific social structural characteristic of our time" (Castells, 2000, 110; quoted in Howard, 2011, 19). He further situates the ascendance of digitally mediated networks historically, as being the result of specific capitalist economic forces working over time to both develop the enabling technologies and to justify the social mode of organization built around those technologies. Rather than attempting to explain social reality as such, Castells's theoretical framework then merely provides a sophisticated language for describing social phenomena at multiple levels of organization and analysis.

The economically-oriented frame and practical orientation of Castells's theory then makes it especially suitable to supporting Bernstein's framework, at least for the present study. Bernstein himself emphasized that the Internet accelerated variation in modes of distribution and thus should be assessed on that basis where possible (Morais, Neves, & Daniels, 2001), and this concern is reflected in Castells's framework. Bernstein also once described Castells's *Information Age* trilogy as a "magnificent achievement" and "one of the most important… modern bibles, that we've got in sociology," suggesting an at least aspirational compatibility between their two projects (Morais, Neves, & Daniels, 2001, 376). I therefore propose to combine Bernstein's framework with Castells's along the following dimensions: pedagogic device, as the ideological, grammatical structuring function of the overall framework, corresponds usefully to Castells's concept of "network power"; pedagogic discourse, to Castells's "culture" in particular networks; and "contexts," to specific sub-networks.

- **Pedagogic Device ↔ Network Power**: Castells defined four modalities of network power, including network power "resulting from the standards required to coordinate social interaction in the networks"; the networked power "of social actors over other social actors in the network"; the network-making power "to program specific networks"; and the networking power " of the actors and organizations included in the networks… over [those] who are not included in these global networks" (2011, 773). Pedagogic device, as the principle which structures pedagogic discourse, corresponds to some combination of network power.

- **Pedagogic Discourse ↔ Culture**: By "culture," Castells means "the relations of production, consumption, power, and experience—along with the information infrastructure that supports these relations" (Howard, 2011, 57). This concept closely resembles Bernstein's definition of pedagogic discourse as simultaneously *defined through* and *structuring of* relations of production and distribution in/of education.
- **Contexts ↔ Sub-Networks**: This term does not have a special significance in Castells's work, but builds on the basic unit of organization in his theory, the network itself. Where contexts could be defined as *the set of actors, activities, and economic relations required to satisfactorily fulfill the required elements of a pedagogic discourse*, they can be registered as sub-networks in a network composed of such actors, activities, and economic relations. That is, by first conceptualizing network nodes/vertices as *actors* (formal and informal institutions), further nodes as *activities* (educational 'products" such as courses, programs, etc.), and network links as the relations of production/distribution between these actors and activities (e.g. that an institution may produce, develop, accredit, fund, etc., a given activity), a network can be constructed in which the selection of a subset of these nodes and links can then be shown to constitute individual contexts.

One final conceptual support is possibly helpful as a means by which to ensure that contexts can indeed be mapped even when—as is to be anticipated by the formal disruptions assumed to characterize the field—the actors or activities involved are not formally defined. That is, New Institutional Economics (NIE) can provide a language for detecting and classifying "institutions and how institutions interact with organizational arrangements" inclusive of formal as well as informal arrangements via its broad definition of institutions as “the written and unwritten rules, norms and constraints that humans devise to reduce uncertainty and control their environment” (Menard & Shirley, 2005, 1). For example, where contexts can be cast as “organizational arrangements” of actors and activities, defined through pedagogic discourse and grounded in relations of production and distribution within the overall field; then the perspective of NIE supports the view that it should be possible to follow this logic of pedagogic discourse, and this grounding in the economic field, to make visible the actors and activities of contexts even when these are informal or even suppressed.

This mapping approach can finally be characterized as committing to a "sociomaterial" reading of media/technology, i.e. as viewing technology as meaningfully structuring sociality in some sense, but also in turn being socially produced. It also commits to the position that the social is meaningfully analyzable using networks, including network science techniques such as measures of centrality and density, while avoiding commitment to any more radical network theory of sociality *per se*. The abstraction and yet historical, technological grounding achieved by this combined "stack" of concepts from Bernstein and Castells then, finally, positions a reading of educational activity that avoids foreclosure around particular historical institutional formations (such as academic institutions), practices (such as instruction methods/paradigms), or environmental conditions (such as classrooms or technologies). Most importantly, it ensures that findings for any one given technology, actor, practice, etc.; or at any level, such as an individual context or a macro-level ideological device; remains comparable to other findings at-level, and translatable across levels.

Summary

This chapter first reviewed academic literature about data science education to show the need for application of a sociology of education lens, not only because this has not yet been applied in the literature, but also because the explicit claims of that literature—and the multivocality with which they are made—speak to a field of some contestation between multiple groups in the academy. In particular, I argued that the especially flexible, multi-dimensional conceptual framework of Basil Bernstein, with its explicit accounting of power and control relations in the "pedagogic discourse" of education, is suited to studying data science education. As part of this, I proposed a method of accounting for a gap in Bernstein's work, as well as a weakness in sociology of education generally, related to the theorization of technology within the field. Namely, I proposed that Manuel Castells's theoretical project support a "mapping" process between Bernstein's and Castells's, then to the broad field of educational activities in the case. The resulting conceptual foundation of the study is network-oriented and concerned with generating empirical insights both at and across levels of analysis. In the following methodology chapter, I detail how this framework and set of values are translated into the research design.

3. Methodology

Introduction

The methodology for this study is exploratory and flexible, using a multiple, nested case study design comprising two levels of analysis and thus two distinct phases of data collection and analysis. These two levels correspond to the lower two layers of the conceptual framework proposed in the literature review chapter: first, a meso-level analysis describing networks of production and distribution in U.S.-based data science education overall; and second, a micro-level analysis of five individual educational contexts.

Following the research problem of finding new ways to study emerging phenomena in the "networked" field of higher education, this chapter also aims to develop a well-integrated conceptual/methodological approach as a contribution to research in its own right. The chapter therefore proceeds by first explaining and motivating the research strategy and design in more detail, in part through an extended discussion of the special role and approach of data collection, modeling, and management in the study. From this foundation, I then discuss the two levels of analysis in more detail, including by presenting a novel conceptual instrument for collecting and analyzing data at the micro/context level.

Research Strategy and Conceptual Structure

The flexible research strategy of the study proceeded from two convictions, first, that previous approaches have been confounded precisely by the closing of conceptual and methodological commitments before the diversity and richness of ground truth could be collected, for example in terms of the narrow framing of technology seen in literature review. The second conviction is inherited from Bernstein, who proposed that research must work from a strong conceptual basis, yet preserve a "conceptual tension which provides the potential for development," that is, must allow continuing refinement of concepts throughout research (Bernstein 2000, 211, quoted in Moore, 2013, 97). The guidelines used for developing the methodology were then taken from Robson and McCartan's definition of flexible research as using multiple qualitative and quantitative techniques to triangulate "truth"; progressive elaboration of design over the course of the study; devotion to a problem rather than a research program*;* and multiple perspectives for conceptualizing research and analyzing data (2016, 147).

From this basis, I chose to apply a multiple case study design for two reasons. First, following Yin's popular formulation, case study is especially fitted to studying "a contemporary phenomenon within its real-life context" for which "boundaries between phenomenon and context are not clearly evident" (2003, 13; quoted in Yaza, 2015, 138). Second, a nested, multiple case study structure fits neatly with the lower two levels of analysis in the Bernsteinian/Castellsian conceptual framework presented in the literature review chapter. Namely, a core case of "the field of data science education" corresponds to a meso-level analysis of pedagogic discourse, realized in terms of the "network of networks" of production and distribution of individual educational activities. Next, by treating multiple individual contexts as smaller cases nested within the larger one, an analytical generalization can be achieved that yet does not reduce contexts to purely quantitative data analysis, i.e. allows the unique texture of each case to work on its own yet contribute to a broader base of analysis.

Data science education itself qualifies as an excellent case for studying the macro-level system of higher education since, from just a cursory review, it presents as not only popular, widespread, and growing, but is also distinctly characterized by a formal innovation of technology and production networks to result in e.g. MOOCs, bootcamps, hybrid degree programs, and more. If the field exhibits a higher-than-normal degree of collaboration between e.g. industry and academia, I view this as beneficial insofar as it at least points to an extreme version of trends that should be detected across other subjects or knowledge fields in higher education. The further specification of U.S.-based education results from a combination of factors: first, based on exploratory data analysis, that most of the actors and activities anticipated by the study exist in the U.S.; second, that the policy context of the U.S. is especially suited to (and positions special stakes for) educational innovation—a theme to which I will return in the discussion chapter; and third, that as a natively U.S. researcher pursuing a qualitatively-oriented study, I am best equipped as a researcher to study this context above others.

Network analysis is critical to the study, not only at a methodological level (as one technique selected from many others), but at a conceptual level, in terms of the importance of networks to the theorization of the field. As such, I treat network data modeling and management itself as part of the research strategy, taking care to translate the conceptual framework into an integrated approach to data collection, management, and analysis that meets several criteria: to natively model data in a network structure, to allow data to be reused across levels of analysis, and to support progressive elaboration of a data model, or ontology, over time. I chose to develop a bespoke database management solution to meet these goals, which I describe in the following section, and document in more technical detail in Appendix: Data Management Technical Reference.

Data Management: Network Data Model and Graph Database Technology

The data management approach is important to the goals of the study in that it translates the conceptual framework identified in the literature review into a formal structure, or system, that yet meets the goals outlined above. Specifically, any system of computational knowledge representation must provide logic, ontology, and computability (Sowa, 2000, xii). Here, the logic is provided by the conceptual framework defined in the literature review, but the distinct challenge at the level of ontology is that, by definition, it is not predetermined ahead of time but is intended to be uncovered by the study itself. At the level of computability, the system must also support strong exploratory flexibility while at the same time protecting data validity and enabling network data analysis. To meet these goals, I chose to use a commercial graph database system called OrientDB to store data, combined with a process for iterating or "refactoring" the data model—that is, the implementation of logic and ontology in the database system—while still preserving data across changes. This approach has been referred to as "agile" or “evolutionary” database design in computing literature, in that, by design, data must maintain “production use” status even as it its structure and contents are continually refined (Ambler, 2003).

The final data model is presented as a class diagram in Figure 1.

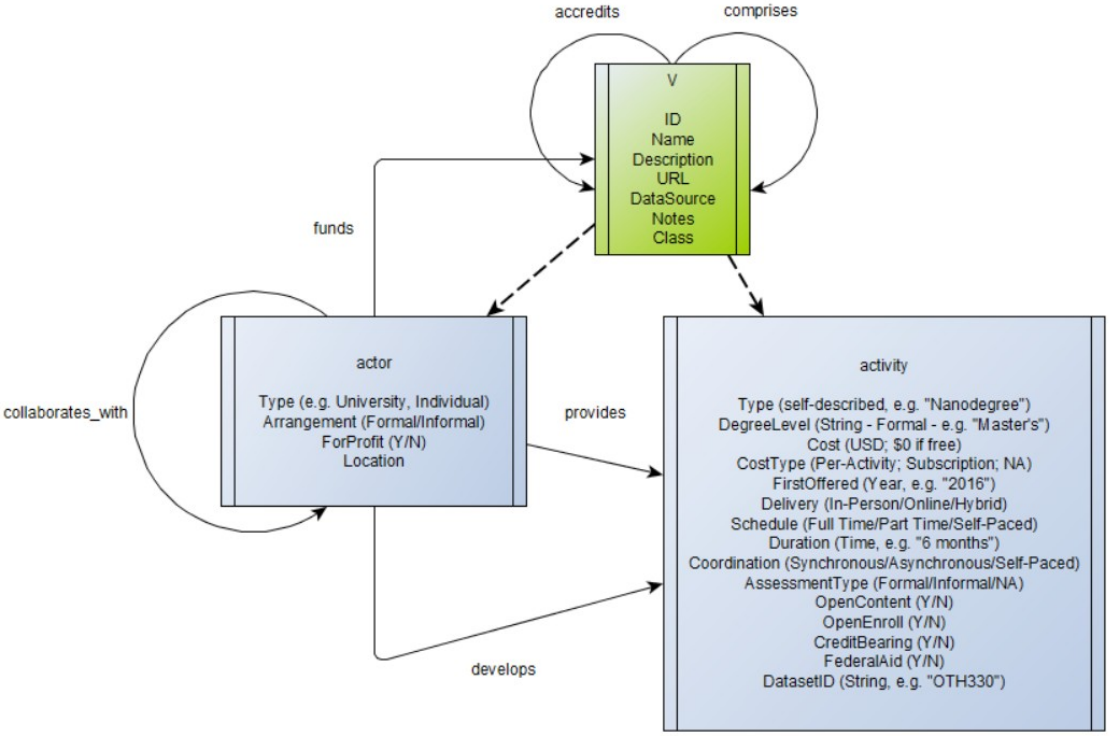

accredits
comprises
V
ID
Name
Description
URL
DataSource
Notes
Class
funds
collaborates_with
actor
Type (e.g. University, Individual)
Arrangement (Formal/Informal)
ForProfit (Y/N)
Location
provides
develops
activity
Type (self-described, e.g. "Nanodegree")
DegreeLevel (String - Formal - e.g. "Master's")
Cost (USD; $0 if free)
CostType (Per-Activity; Subscription; NA)
FirstOffered (Year, e.g. "2016")
Delivery (In-Person/Online/Hybrid)
Schedule (Full Time/Part Time/Self-Paced)
Duration (Time, e.g. "6 months")
Coordination (Synchronous/Asynchronous/Self-Paced)
AssessmentType (Formal/Informal/NA)
OpenContent (Y/N)
OpenEnroll (Y/N)
CreditBearing (Y/N)
FederalAid (Y/N)
DatasetID (String, e.g. "OTH330")

Figure 1: Final data model

Originally composed of 6 distinct vertex types and multiple additional edges, this final model expresses an analytical focus that developed throughout the study to focus especially on the relations of production and distribution within the network proper, and to push additional criteria such as “Delivery Method” or “Certification” into metadata (i.e., properties) on vertices instead. These properties are defined in full in the Appendix. In one example, where multiple courses presented on the MOOC platform Coursera are bundled into a “Specialization”: each course would be modeled as an activity, the specialization would be modeled as another activity with a "comprises" edge linking it to each individual course, and finally the course and specialization activities would be linked via a "provides" link to an actor representing Coursera. Another actor describing the organization that developed the courses (such as a university), would be linked via the "develops" link to each course and to the specialization. OrientDB's implementation of a customized version of Structured Query Language (SQL), a domain-specific language for querying and manipulating data, then enabled me to project subsets of the network in multiple ways according to different analytical requirements, as discussed in the findings/analysis chapter. For example, I could easily view a sub-network of only "courses" and "universities," or only activities that were credit-bearing, or only actors that had collaborated in the production of an activity together.

The use of a graph database is therefore central to the study's conceptual approach; however, some further distinctions are helpful in clarifying precisely how. Barabási, a pioneer of the field of network science, notes that although the terms "graph" and "network" are used somewhat interchangeably to refer to data representation structures defined by nodes/vertices and the links/edges that connect them, the term "graph" technically refers to the mathematical representation of these structures and "network" to their semanticization (2014, 6). Here, OrientDB implements the networks of production and distribution of data science education as a *multidimensional property multigraph*. The unique features of this graph type are as follows: it is *multidimensional* in that it models multiple types of entities or domains, namely "actors" (organizations and institutions) and "activities" (the educational products and services of these actors, such as degrees, courses, etc.); it is a *property graph* because in addition to modeling vertices and edges, it stores metadata directly on vertices and edges which describe them (in this case, for example, whether an actor is for-profit, or whether an activity is credit-bearing); and finally, the graph is a *multigraph* because its edges can connect not only between two vertices of the same type, but also between two vertices of different types (actors and activities) as well as to and from the same vertex (as in the "comprises" relationship). Together, these features uniquely support the research goal of.

In short, the graph database management system enables a flexible, integrated, expressive, and progressively elaborated approach to collecting and modeling data, per the goals of the study, by managing research data as a multidimensional property graph, and by providing streamlined data manipulation capabilities, such as SQL, for interactively exploring and refining the data and data model throughout collection and analysis.

Data Collection: Strategy, Techniques, Limitations

Data was initially seeded from a variety of bulk sources, notably a dataset provided early in the study's planning process by Class Central, a website devoted to tracking MOOC activity since 2013. I later received a bulk dataset from Institute for Advanced Analytics, a research center at North Carolina State University. These initial imports of tabular, historical data entailed translating source data structures into a provisional metadata data model. Exploratory visualization and analysis of this data then helped me refine the metadata structure finally presented in Figure 1. I preserved provenance throughout the study using the "DataSource" property seen in Figure 1, to support data quality as well as to ensure that any future data sharing efforts could be conducted in accordance with the wishes of data providers.

After this initial phase was complete, I applied multiple techniques to collect and code additional data. This included manually performing qualitative review of individual web pages discovered through review of bulk data gathered from the above sources, from structured web queries, and from my own archive of bookmarks data science education programs. This qualitative review entailed translating publicly accessible content published on websites into the metadata structure shown in Figure 1. I also journaled throughout this process, using the journaling process as well as exploratory analysis in graph database visualization tools to help critically evaluate the data model, refine search methods, adjust search strategy, and maintain alignment to the research problem of uncovering new institutional arrangements and pedagogic discourse within the field. Technically, I performed these tasks using a combination of tools for web scraping, data preparation and integration, and visualization. Such techniques are generally commonplace in mixed-method research, where for example the use of a research diary/journal and visualization techniques are often used to facilitate the quality of research (Tracy, 2010; 2012; Creswell, 2003), even if their organization into the database system described above is somewhat more unique to this study's approach.

In terms of the ethics and values of this data collection process: I reviewed hundreds of individual pages of content, including individual classes, but always in a manner consistent with relevant terms of service. For example, I did not include data that would be considered proprietary to a private signup/membership/subscription, but rather limited myself to accessing and collecting only what was available on the open web. I did repeatedly use the Internet Archive Wayback Machine (https://archive.org/web/) to access content that had been retired, moved, or altered, but I view this as consistent with the overall approach. Following, I note some additional limitations and delimitations of data collection in the study.

Limitations

- The inclusion of data from Class Central and similar catalog/index websites results in an uneven snapshot of historical information. I consider this consistent with the goal of surfacing more unique patterns in the field, rather than capturing a comprehensive snapshot of a single point in time or of tracing specific historical developments over time.

- I rely solely on information presented on the open web. I did allow myself to utilize archived web data, for example, but did not use e.g. interviews or other methods of gaining insight beyond what is available to a non-logged-in web user.

Delimitations

- Activities must position themselves as formal education through rhetorical cues such as the terms “course,” “degree,” “program,” etc., the use of assessments such as quizzes or projects, etc. Bernstein's own test that education must operationalize curriculum, pedagogy, and assessment will in turn be applied, not to data collection, but data analysis.
- Activities and organizations must explicitly fashion themselves to be related to “data science” or “data scientist” education. I operationalized this by requiring that the string “data sci” (such as “data scientist” or “data science”) appear directly in the text of either the title, description, or related marketing material for activities/organizations. In cases where an activity is listed in a broad category, such as in the Class Central dataset or in the catalog of a MOOC platform, but that activity did not otherwise seem to consider itself dedicated to data science, I did not include it in the final dataset.
- Organizations or activities must be based in the United States, or sold/delivered on platforms that are based in the United States. For example, I included a course offered by a French university on the American MOOC platform edX, and so therefore also included the French university itself.

To reiterate, these choices are compatible with the goals of the study as exploratory, i.e., as designed to surface and analyze novel institutional configurations for their pedagogic discourse rather than to provide a comprehensive economic or strategic analysis of the field *per se*.

With overall approach, a database management system and initial data model, and data collection rules in place, the meso-level methodology can be explained.

Research Design: Meso-Level - Network Analysis

The primary goal of this level of analysis was to map the field of U.S.-based data science education in such a way as to uncover patterns of institutional arrangements within it. By modeling collected data directly as a network of "actors" and "activities," with multiple relationship-types describing how they relate, I could then use the ad-hoc querying capability of OrientDB, combined with the basic network statistics provided by Gephi, a popular graph visualization and analysis tool, to explore different perspectives and scenarios in the data computationally. This process effectively uncovered additional, connected actors and activities organically while also allowing me to develop an initial sense of specific patterns or cases that might be interesting to explore in more detail at the micro level. This secondary goal, of driving case selection at the micro level, was then refined as the dataset developed using network visualizations in OrientDB and Gephi. Additional data collection, as discussed above, involved conducting additional web searches.

An important conceptual and methodological choice which arose at this level was how to conduct network analysis. Network analysis provides simple descriptive techniques, such as measures of degree (number of edges connected to each vertex), as well as predictive or classificatory algorithms, such as community detection (Barabási, 2014). The multidimensional/

multilayered quality of the graph data complicates the application of simple and especially more complex techniques (Kivelä et al., 2014), while on the other hand, the much more descriptive and qualitative orientation of the approach used by Castells—see especially Arsenault & Castells, 2008, 713—is not suited to the exploratory breadth of analysis pursued by this study. I chose to balance these two extremes by creating multiple visualizations for qualitative interpretation, then, creating a specialized subset of the network known as a collaboration network, to perform more simplified (that is, not "multilayer") analytics on. In support of this, I relied on the querying and filtering capabilities of OrientDB to subset and experiment with multiple way so visualizing and analyzing data. To create the collaboration network, for example, I subsetted data by traversing the graph to discover all instances where two actors were connected to a single activity, then storing that information back to a new "collaborates_with" relationship between the two actors, until all such pair-wise combinations of actor collaborations were found. In short, I prioritized a more qualitative approach to analysis, but leveraged computational aspects of network data analysis, including calculation of simple network statistics like degree, to support and guide this analysis.

Whereas meso-level network analysis was intended to assess the overall field and to aid in case selection, analyzing pedagogic discourse involves tracing its operation at the scale of individual contexts. This was performed in the next phase in a series of five smaller case studies.

Research Design: Micro-Level - Case Study

The primary goal of conducting this second, lower-level case analysis was to support stronger analytical generalization overall by enabling individual contexts to be evaluated on their own terms, thus highlighting distinct configurations of pedagogic discourse on a per-context basis. The section's secondary goal was to test and refine the conceptual apparatus developed in the previous chapter, especially in terms of the performance of an additional conceptual instrument created to help bridge Bernstein's framework to the networked field of education considered by the study. This instrument is described in more detail below.

Procedurally, I first selected cases that I felt were representative of a variety of types of educational activity, institutional type, and pattern of network production. This was informed by the analysis performed at the meso-level. I then reviewed each case using a common template as a starting point, finally allowing distinct themes to take precedence within each case summary at time of final write-up for the findings/analysis chapter. I present the template in Figure 2.

- Evidence
    - Context network diagram
    - Autonomy/Rationalization (A/R) graph
    - Classification & framing rule: Hierarchy
    - Classification & framing rule: Sequencing
    - Classification & framing rule: Evaluation
    - Site/Source of knowledge production
    - Knowledge structure orientation (hierarchical vs. segmented)
- Themes
    - Power and control relations
    - Visible vs. invisible pedagogy
    - "Powerful" knowledge: how is knowledge framed, justified, translated?
    - "Location" of pedagogic device

Figure 2: Case study template

Conceptually, the approach is to deemphasize educational technology as a defining or framing concept and instead focus on the functional definition of educational contexts in terms of classification and framing rules, participating actors, etc., as outlined in the literature review chapter. I additionally wanted to take notice of specific themes which connected back to the research problems that originated the study, e.g. what power and control relations exist in data science education. The template was therefore designed to guide the capture of evidence for the same dimensions in each context, aiding in comparability across cases even where the case study approach does not support statistical analysis (Robson & McCartan, 2016, 154).

As part of this overall template, I also chose to create unique instrumentation to help account for the special economic dimension that the conceptual framework took on in the literature review chapter. For example, despite the importance of economic concepts which I argued for above, Bernstein did not directly discuss economics *per se*, and so there is no dedicated language in his framework to describe it. I therefore conceptualized a graphical representation of the relative economic autonomy and/or rationalization at play in a given context to help provide such a language. I turn now to explaining this graph in more detail.

The Autonomy/Rationalization Graph

I propose the Autonomy/Rationalization (A/R) graph as a conceptual instrument for describing context networks in terms of the balance of economic “pressure” to which the educational activities within that context are exposed. By considering all formal educational activity to exist in some kind of market (in the broadest sense, as always representing one option among many others and therefore competing for access to limited resources), then every educational context should demonstrate a unique balance of economic “pressure” across different activities within the context. In practice, this generally manifests as the relative market exposure of educational activities at various levels within the context, for example by a course versus a degree program, but it could also be applied to formally non-market contexts by applying a New Institutional Economics lens. This is an important dimension to consider for two reasons: first, that whereas this A/R dimension would have been more fixed in a mode of traditional higher education practices, disruption of this dimension is now one of the distinguishing differences of the new field; and second, that these different economic configurations should be anticipated to have some influence on a context’s pedagogic discourse, so that first making that configuration more visible and comparable across contexts becomes important to aid in analysis.

I propose the graph be generated as follows. For a given context-network, and at each of four generalized, hierarchical levels of educational activity within that context, each activity level should be rated along a standardized axis describing the exposure of that level to economic competition/marketization, where both levels were calibrated through exploratory data analysis in this study. I propose these activity levels to be classified as *tutorial* (in the sense of 1:1 teaching relationships), *lecture/lesson*, *course*, *curriculum*, and *catalog.* For any given context, some activity can occur at any of these levels, such that higher levels integrate activity at lower levels. I next propose that the levels of economic competition/marketization be classified as 3 - *exclusive*, 2 - *optional*, 1 - *submerged*, and 0 - *absent* exposure. For any given context, translating these characteristics into a vector space of activity level vs. rationalization level then produces a chartable line which expresses the relative level and spread of economic rationalization across the spectrum of activities within the context.

Two examples help illustrate this model. First, a hypothetical case representing an extreme stereotype of the "national-liberal" university can be imagined, whose graph appears as follows:

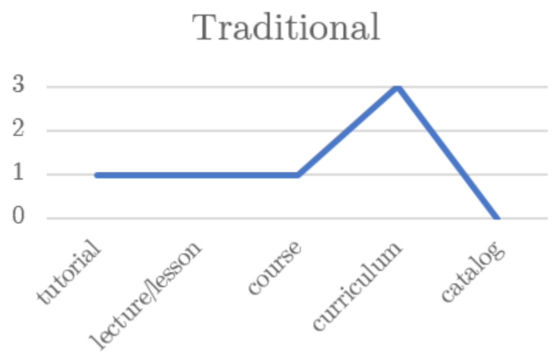


Figure 3: A/R graph: Traditional context example

In this example, the university receives state funding to administer degree programs (curricula) in a marketplace of other universities doing the same, such that courses, individual lectures, and tutorials are subsumed under the degree/curriculum with respect to market exposure. This indicates a high level of autonomy on the part of the university overall and a low level of rationalization with respect to market forces.

In contrast, a fully rationalized educational system could be hypothesized in which the university is displaced (or replaced) by a marketplace in which education is coordinated at the highest possible level of rationalization, namely the tutorial. This would obtain the following graph:

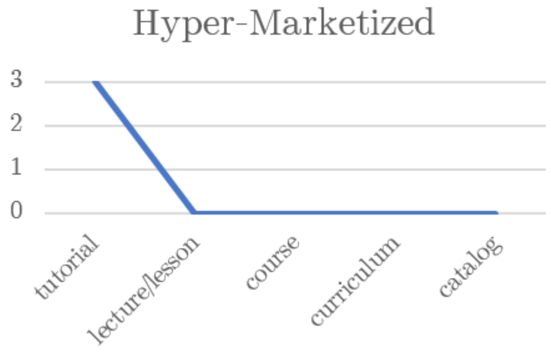


Figure 4: A/R graph: Hyper-marketized context example

At these two extremes, regardless of other pedagogic-discursive rules such as hierarchy, sequence, and evaluation; or yet regardless of stated political orientation, such as progressive versus traditional; what should be apparent is that the economic pressures at work in each context can yet be vastly different, per the shape of this graph, with the question following of *how* these forces ultimately impact pedagogic discourse. I take this question up as part of the findings/analysis chapter.

In summary, the strength of this conceptual tool is that, like the other concepts in Bernstein's conceptual framework with which it is designed to operate, it is general and generalizable; for example, it does not pre-assign political meaning to any particular configuration, and it can be applied to any context regardless of the formality of its economic coordination or its materiality/ use of technology.

Summary

Finally, the methodology of this study is decidedly “Bernsteinian,” in that it prescribes a flexible, case-oriented approach, not only due to conditions found in the arena of study which require exploration and description given the current state of literature regarding it, but also due to an additional, intrinsic goal to refine conceptual modeling and instrumentation which, when connected to a broader project in the sociology of knowledge, point the way toward future study. In the following chapter, I combine presentation of findings and analysis.

4. Findings/Analysis

Network Analysis: Patterns of Economic Coordination

In total, I collected 915 educational activities and 330 actors, totaling 1,245 vertices and 2,755 edges. By viewing a network visualization of all actors and activities (see Figure 5), a few patterns emerge. First, there are two large networks of densely inter-connected actors which account for over half of the dataset, then, many much smaller networks. Notably, the mix of *types* of actors are not too different between these three network scales: university/degree pairings make up most of the less-connected pairs shown around the outer ring of the visualization, but universities, course platforms/marketplaces, and other individual actors do appear across all network types.

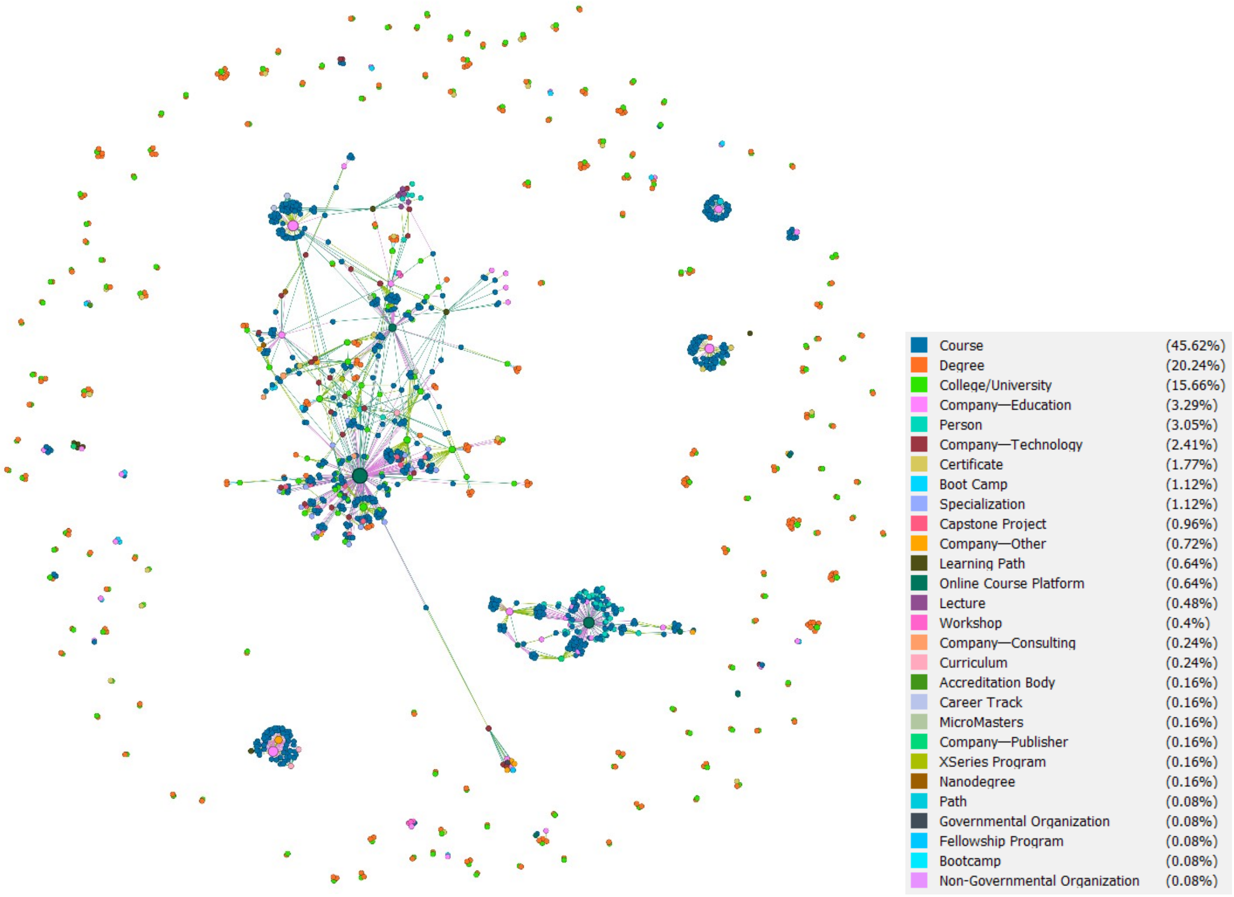


Figure 5: Complete meso-level network

The inclusion of activities however not only clutters the graph, but also artificially increases its density in an uneven way, for example where a complete set of courses in a program is modeled rather than solely the program itself. By removing the activity nodes and focusing instead only on collaborations between actors, a more intuitively useful graph can be discerned. Here, wherever any two actors are both linked by any kind of edge to the same activity, then these two actors are considered "collaborators." This (Figure 6) is the "collaboration network" referred to in the previous chapter, shown here with text labels stating the name of the actor.

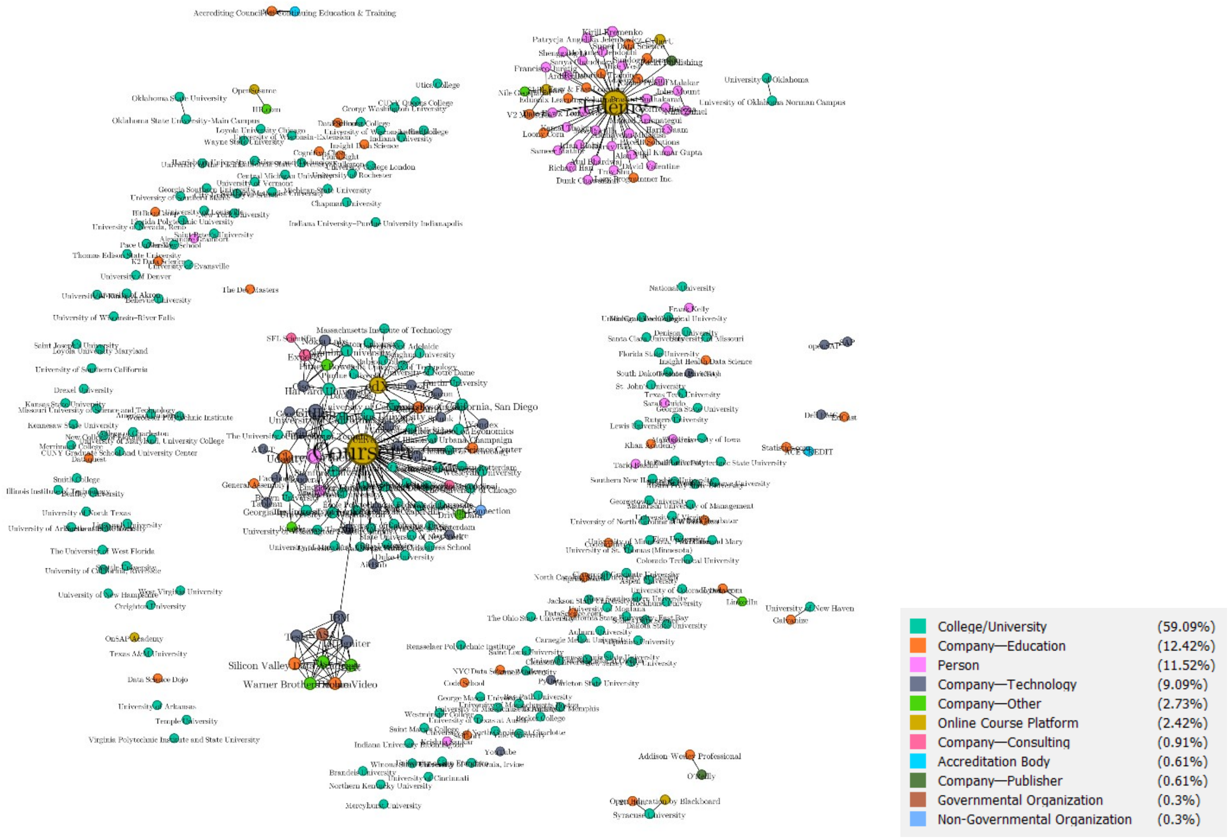

College/University (59.09%)
Company—Education (12.42%)
Person (11.52%)
Company—Technology (9.09%)
Company—Other (2.73%)
Online Course Platform (2.42%)
Company—Consulting (0.91%)
Accreditation Body (0.61%)
Company—Publisher (0.61%)
Governmental Organization (0.3%)
Non-Governmental Organization (0.3%)

Figure 6: Collaboration network

Where questions of link "strength," for example collaboration across multiple nodes, would be artificially increased by the multi-level and nested nature of some (but, crucially, not all) of the activity data, a simple, unweighted edge approach was taken here. In practice, an example of where this became important was for in Coursera's "Specializations," a term used by Coursera to refer to bundles of courses that result in a specialized certificate (in addition to individual course certificates) and sometimes features a Specialization-exclusive "Capstone Project" course. Many of these are advertised, at the Specialization level, as being co-developed with companies, especially technology companies like Microsoft. While listed at a Specialization level, however, these relationships are not always detailed at the level of individual courses. By taking an unweighted approach, the collaboration between Coursera, Microsoft (in this example), and a university can become discernible, without being artificially lowered in strength in comparison to another case where the relationship may instead be clearly defined for each course.

By filtering out vertices that lack any collaboration whatsoever, only 155 actors remain—just under 50% of all actors in the network. Put another way, about half of the activity discovered by the study was produced independently with respect to the overall field of actors involved in producing data science education, while the other half worked in partnership with at least one other organization. These remaining networks can be explored through some additional visualizations and queries. For example, the overall graph (above) shows a few core networks, one around a highly connected set centered around the MOOC platforms Coursera and edX, and another centered around Udemy, a Self-Paced Online Course (SPOC) SPOC marketplace. Notably, the Udemy network (which also contains competing platforms Skillshare and CyberU) does not contain any university actors, but rather contains solely commercially-produced educational content.

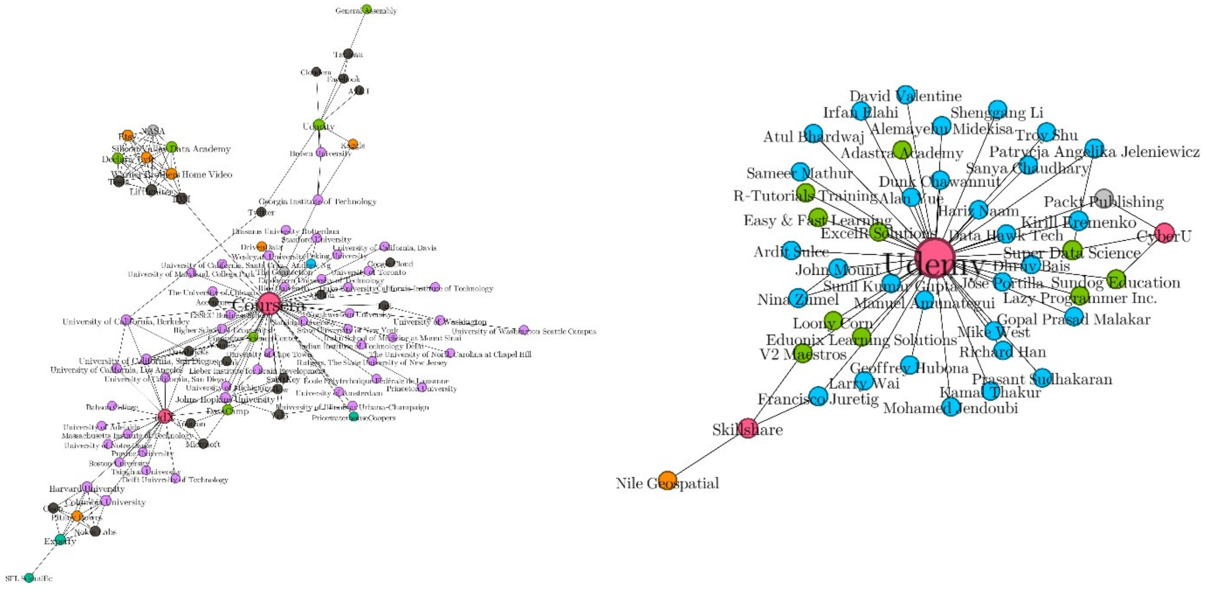


Figure 7: Two largest collaboration sub-networks

This lead me to question what the collaboration graph would look like without the involvement of the super-connected nodes of edX, Coursera, and Udemy. Conceptually, this means separating out marketplaces to focus on more development-oriented collaborations. The result, shown in Figure 8, displays only vertices having at least one collaboration, after removing edX, Coursera, and Udemy:

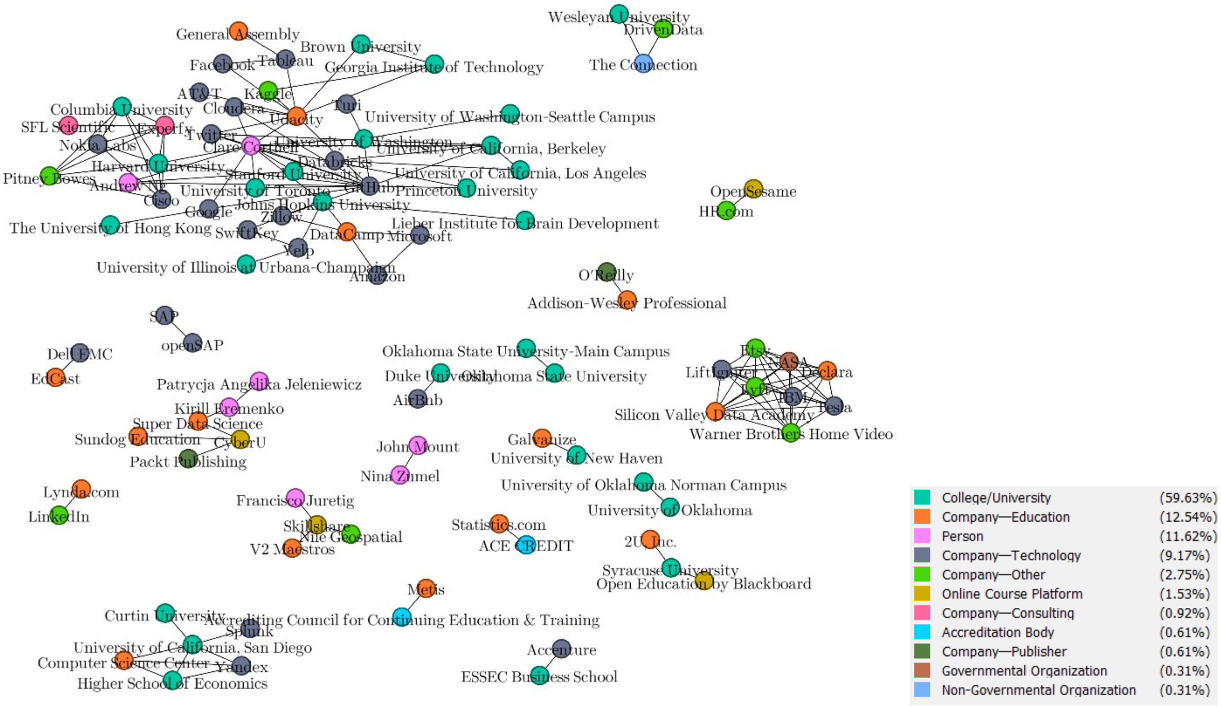


Figure 8: Meso-level collaboration sub-networks without super-connected actors

The resulting network of 104 "connected" or "collaborating" actors is much less densely connected than in other visualizations, but a large sub-network still appears which features multiple universities, technology companies, and education companies. In total, 47 of these 104 actors (about 45%) are connected to each other via this sub-network, which has an average degree of four. If edX and Coursera are added back in, the resulting sub-network maintains an average degree of about 4, and comprises 99 actors, or is 30% of *all* actors and about 60% of all *connected* actors.

Even given the limitations noted in the preceding chapter, this result points to the field being not only relatively well-connected and collaborative, but connected and collaborative *inclusive of academic as well as non-academic actors,* such as educational companies and technology companies. However, although the mix of *types* of collaborative institutions is relatively diverse, there is less diversity in the relative status of those institutions which are collaborate more. This can be seen more clearly in Table 1, which subsets only U.S.-based institutions from the densely connected sub-network, showing recent national ranking data for each institution. Notably, out of thousands of colleges and universities in the U.S., all the ones in the core network rank within the top 100, with 5-6 out of the top 10 for each ranking list appearing in the network.

University	**ARWU**	**Forbes**	**U.S. News & World Report**	***Average***
Brown University	45	8	14	*22*
Columbia University	7	16	4	*9*
Georgia Institute of Technology	47	89	34	*57*
Harvard University	1	4	2	*2*
Johns Hopkins University	14	66	10	*30*
Massachusetts Institute of Technology	4	**5**	7	*5*
Princeton University	5	**3**	1	*3*
Stanford University	2	**1**	5	*3*
University of California, Berkeley	3	40	20	*21*
University of California, Los Angeles	10	46	24	*27*
University of Illinois at Urbana-Champaign	23	72	44	*46*
University of Michigan	17	44	27	*29*
University of Texas at Austin	30	93	56	*60*
University of Washington	13	75	54	*47*

Table 1: Ranking data for highly connected universities; source: Wikipedia

Notably, Bernstein actually hypothesized that in emerging practices of using "hypertext" and the Internet, it would be non-elite universities which would disproportionately be forced to use new discourses, whereas elite universities would be slower to adapt (Morais, Neves, & Daniels, 2001). Instead, the collaboration between *more* rather than *less* elite institutions suggests a power-enhancing quality to the formal innovation pursued by these actors, rather than a reactive, compensatory one as suggested by Bernstein. In Castellsian terms, this might be understood as an example of *network-making* power on the part of platforms like edX and Coursera, *networked power* on the part of elite universities, *network power* with respect to the pedagogic-discursive influence wielded by this densely connected network, and thus *networking power* over those institutions which are not included in the network (Castells, 2011, 773). This might in turn constitute the pedagogic device of a dominant thread, or paradigm, of data science education in the U.S. Testing this last claim, however, not to mention uncovering evidence of alternative discourses, device, or "counter-power" in the overall network, is the domain of micro-level case/ context analysis.

Case Analysis: Networked Education Contexts

Overview

As discussed in the methodology chapter, the following five cases were selected to represent a diverse mix of institutional and educational activity types, as discovered during meso-level analysis. Detailed in Table 2 using data collected during meso-level research, these are: two contexts centered around universities, one from without and another from within the "elite" sub-network described above; two contexts centered around activities that overlap with university activities but are not directed by them *per se*, each featuring a distinct pedagogic and economic approach; and one from within the completely non-university-affiliated sub-network.

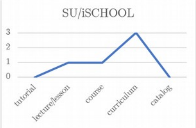
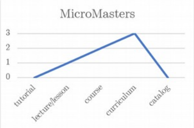
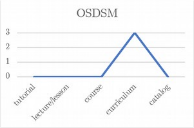
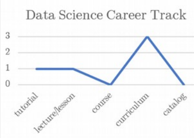
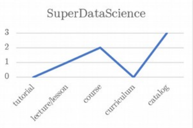
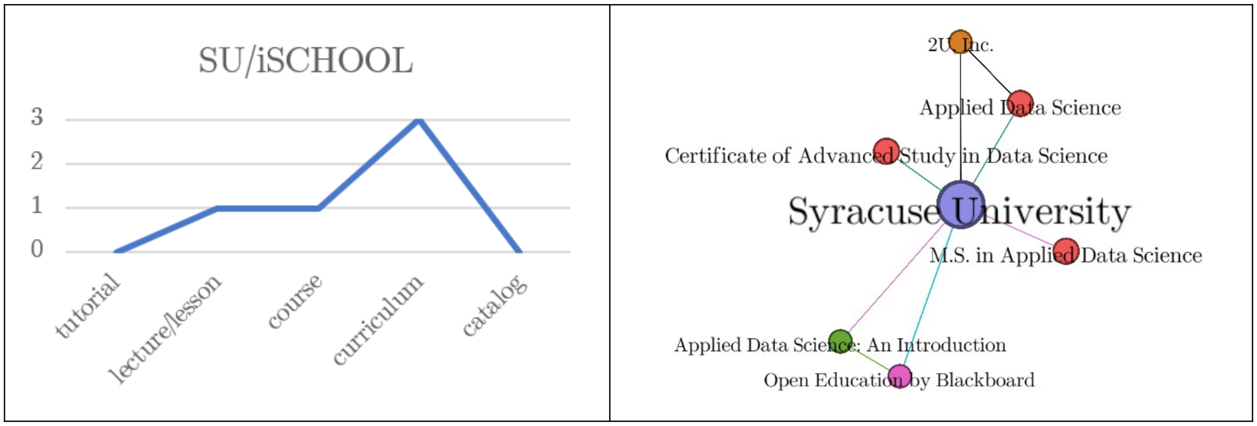
Context Cases: Overview and Summary					
Name	Master of Science in Applied Data Science	MicroMasters in Data Science	Open Source Data Science Master's	Data Science Career Track	SuperDataScience
Provided By	Syracuse University; 2U, Inc.	edX	GitHub	Springboard	SuperDataScience
Developed By	Syracuse University; 2U, Inc.	University of California, San Diego (UCSD)	Clare Corthell	Springboard	SuperDataScience
Accredited By	Syracuse University	Curtin University	N/A	N/A	N/A
A/R Graph					
Type	Master's Program	MicroMasters™	Curriculum	Workshop	Subscription
Cost (USD)	$54,000	$1,400	N/A	$4,800 *or* $1,000/ month	$35/month
Delivery	On-Campus or Online	Fully Online	Fully Online	Fully Online	Fully Online
Scheduling	Full Time; Synchronous/ Asynchronous	Part Time; Asynchronous	Self-Paced	Self-Paced	Self-Paced
Assessment	Formal	Formal	Informal	Informal	Informal
Eligibility	Application required; limited availability	Open enrollment; account required	No enrollment required; no account required	Application required; limited availability	Open enrollment; account required

Table 2: Overview and summary of micro-level context cases

The following case-specific sections focus on the economic and pedagogic features of each context in more detail.

Case 1: Master of Science in Applied Data Science (Syracuse University; 2U, Inc.)

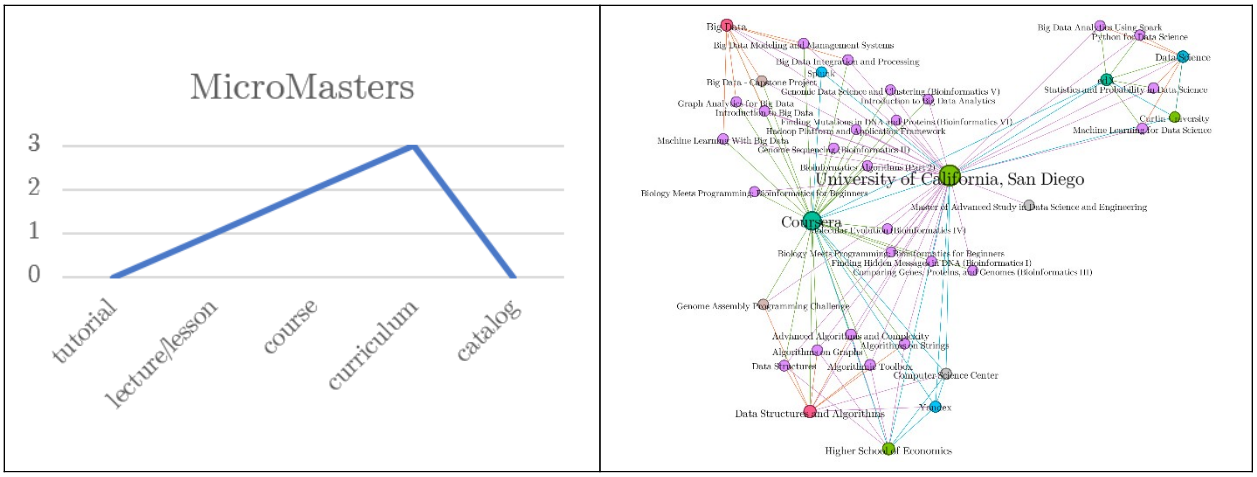


Syracuse University (SU) offers a Master of Science in Applied Data Science, starting 2017, in both a campus-based and online variant; they've also published on the subject of data science education (Stanton, et al., 2012) and offered other forms of training since at least 2013. In terms of selection and sequencing, these new programs were developed through partnership between the university's management school and School of Information Studies from content originally developed for previously taught individual courses. The curriculum emphasizes accessibility for non-technical audiences, with no required courses in math or statistics, though these are offered as electives. In terms of evaluation, the programs culminate in a faculty-reviewed project portfolio requirement rather than exams or a thesis. Overall, the program then seems to conceptualize data science as a “segmented” but “integrated” knowledge structure, in Bernstein’s terms; they also depict an “invisible pedagogy,” i.e. rather than assessing students against explicitly defined technical performance criteria, the program conditions students to perform against unspoken or incompletely elaborated social criteria, such as the adoption of certain cultural values. This is consistent with the applied focus of the degree as well as the program's primary site of operation within the iSchool. More interesting is the difference between the two programs introduced at the level of hierarchical rules by the online context.

This difference begins with the separation into distinct websites for each program. While the website for the on-campus program advertises the online variant, and claims it to be “just in a different delivery mode,” the online program makes no mention of the on-campus program. Rather, the online program emphasizes benefits added by a “program partner,” 2U, Inc., who provide a proprietary web-based and mobile software platform, as well as third party administrative staff, including admissions, career counseling, and IT support. 2U's role is characterized as enabling an innovative, interactive access to other students as well as to faculty, i.e. as a way of reproducing the hierarchical classification rules of a traditional on-campus experience.

This hierarchical orientation has two key impacts for the case. First, the structuring of access to faculty and separation into access to unique staff suggests that the online degree might qualify as a distinct context, or might not, depending on the extent to which the promise of providing students with similar access to faculty and to each other is reproduced online. Here, curiously, the fact that online students have access to different supporting staff is a notable economic innovation but possibly not a pedagogic one. Second, the appearance of 2U, Inc. in the production network mean that, even if not intended, the online program is likely to have a conditioning or disciplining effect on the on-campus program over time; that is, the on-campus program is pedagogically entangled in the online program, and vice versa; so that whether this is positive or negative may in turn depend on the extent to which—according to the first impact—the online context is more or less distinct from the on-campus one.

Case 2: MicroMasters in Data Science (University of California, San Diego; edX; Curtin University)



The term "MicroMasters," a trademark of edX, refers to "a series of graduate level courses from top universities… Students may apply to the university offering credit for the MicroMasters certificate and, if accepted, can pursue an accelerated and less expensive Master's Degree" (edX Inc., 2016). Compared to the "Nanodegree" from Udacity, "Specializations" from Coursera, or

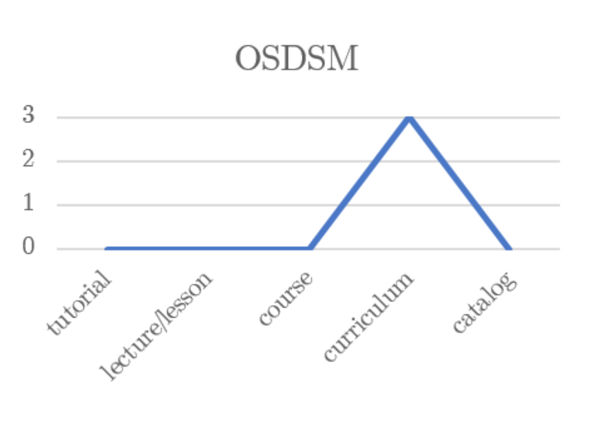
“XSeries” and “Professional Certificate” programs from edX, the MicroMasters is unique as a bundling technique in its funneling of students into an on-campus program. Curiously in this case, that on-campus program is not from UCSD.

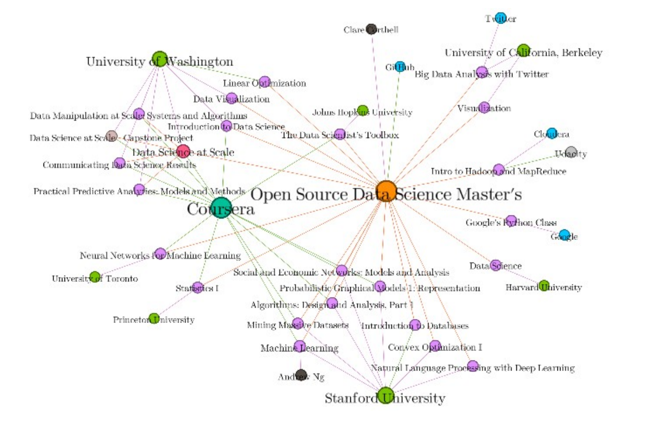
UCSD has a large network overall, including multiple courses and programs on Coursera, on campus, and through collaborations with technology companies as well as other universities. The MicroMasters does not, however, connect to or discuss any of these other activities, nor do these mention the MicroMasters; rather, the MicroMasters enforces strong classification boundaries between it and other data science education offerings. Strong boundaries are also enforced between elements within the program, such as through separations between free (“Audited”) and paid (“Verified”) tracks, between interactive elements on a per-activity basis, and even between cohorts of users engaged in the course over time. All content is technically free to access, for example, but requires registration, and additionally, payment, to qualify students for premium support and formal certification at the end of the program. The courses are also intended to be taken sequentially, with “previous courses in the MicroMasters program” listed as prerequisites in some cases. One way of interpreting this is that only those students with the luxury of committing to a year of study will be able to complete the MicroMasters, diminishing some of the benefit of flexible time access. Course reviews indeed cite this pacing strategy as a reason for choosing other programs that take less time.

Strong boundaries also appear to exist between actors. The program that the MicroMasters funnels into is offered by Curtin University, an Australian university that does not otherwise appear to collaborate with UCSD in any data science endeavors. Curtin University also doesn't appear to be involved in any way with the development or operation of the MicroMasters, but rather are listed only because they accept the courses toward credit in their own full-time, on-campus Master's program. Curtin University also offers an on-campus BSc in Data Science as well as two of their own MicroMasters programs on edX, but they do not, except for their listing on the UCSD MicroMasters webpage, otherwise mention their involvement in the UCSD MicroMasters across other data science offerings or their main website. Given the limited benefit of credits for a two-year, on-campus, Australian degree program to students based anywhere besides Australia, their inclusion seems designed, at least in part, merely to qualify the MicroMasters as such.

The positioning of this context within UCSD's larger network, including the participation of Curtin University; and its lack of mention by Curtin University or UCSD; finally positions the discourse of the MicroMasters as driven by a strategic partnership interest among the parties, i.e. to produce a MOOC product. The edX platform's affordances, institutional partnerships, and economic interests may therefore be seen to play a discursive role, exerting a "networking power" in bringing together other institutions to create an educational "product" which, by legitimation through these institutions, serves to reinforce edX's power in the field.

Case 3: Open Source Data Science Master's




Clare Corthell
Twitter
University of Washington
University of California, Berkeley
GitHub
Big Data Analysis with Twitter
Linear Optimization
Data Visualization
Johns Hopkins University
Visualization
Data Manipulation at Scale: Systems and Algorithms
Introduction to Data Science
The Data Scientist's Toolbox
Data Science at Scale - Capstone Project
Cloudera
Data Science at Scale
Udacity
Communicating Data Science Results
Intro to Hadoop and MapReduce
Practical Predictive Analytics: Models and Methods
Open Source Data Science Master's
Coursera
Google's Python Class
Google
Neural Networks for Machine Learning
Social and Economic Networks: Models and Analysis
Data Science
University of Toronto
Statistics I
Probabilistic Graphical Models 1: Representation
Harvard University
Algorithms: Design and Analysis, Part 1
Princeton University
Mining Massive Datasets
Introduction to Databases
Machine Learning
Convex Optimization I
Natural Language Processing with Deep Learning
Andrew Ng
Stanford University

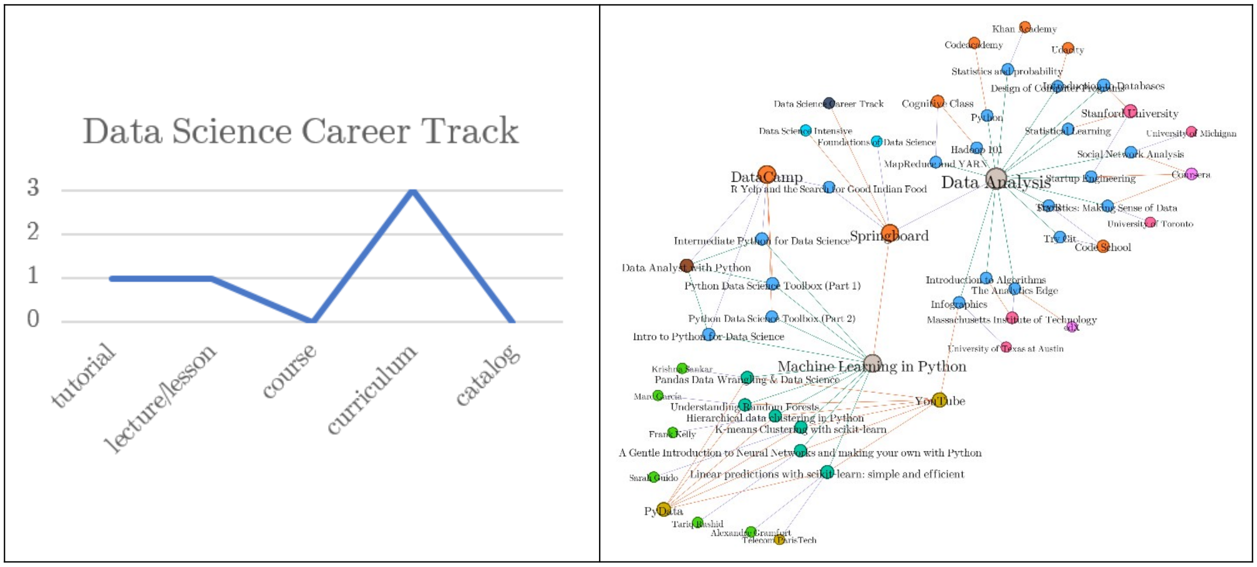
The Open Source Data Science Master's (OSDSM) is an "open-source curriculum for learning Data Science" that "breaks down the core competencies necessary to making use of data"

(Corthell & @datasciencemasters, 2017). It consists of a set of files, stored on GitHub, which structure annotated links to courses, books, tutorials, and other resources. After being created by a single individual in 2013, the OSDSM has had 294 edits contributed from 36 contributors, resulting in 7500+ “stars” (essentially, bookmarks) in the GitHub user community. The technical affordances of GitHub, ostensibly designed for managing open source software, are thus repurposed by the OSDSM community to coordinate maintenance and improvement of the curriculum. For example, a GitHub “issue”—a discussion feature used for tracking software bugs—was opened by users to note that a course in the curriculum had been taken offline. Other users then provided backup copies of the course on filesharing networks, closed the issue, and updated the curriculum to point to the issue history instead of the broken link. Not all links are kept up to date in this way, but the potential for them to be updated remains.

Translated into the A/R graph, this reliance on volunteer labor suggests that even though no money need be paid by transmitter or acquirer (GitHub hosting is free, as is almost all curriculum content), a limited resources must still be spent to maintain the OSDSM, namely creative labor and time. This volunteer labor should also be understood as serving a pedagogic function, implementing assessment and evaluation: where other programs emphasize the need to coordinate projects, for example, this curriculum, by merit of its presentation as an open source project, suggests that users use the curriculum itself as a record or project-based example of their progress as such. Indeed, the curriculum has been “forked” on GitHub—a unique copy instantiated in the files of individual users, for their local editing—3700 times. This forking activity then performs multiple functions simultaneously: it provides the basis for students to track progress; ensures, and proves, that students are technically and socially capable of working like data scientists; improves the performance metrics and thus status of the activity overall; and blurs the line between students and contributors.

Pedagogically, although the OSDSM applies a traditional sequencing orientation, its hierarchical structure—with this open invitation to collaborate—presents a double meaning in terms of power and control relations. Ostensibly empowering, it also conceals a discursive claim about what it means to be a data scientist, namely that data scientists must not only possess the skill and discipline to learn complex educational content alone, but must also be able and willing to use GitHub to access, track, and co-construct that content—not only unpaid, but without access to financial aid of any kind. This discursive orientation is not incompatible with the creator's clear commitment to the ethical application of data science and the democratization of access to its education, but rather expresses an underlying discourses of lifelong learning and creative labor which attend the elite intellectual class of which she is a member; namely, that workers in the "new" or "data" economy on some level compete on the basis of their knowledge and can therefore be expected to self-organize the educational activity required to develop themselves.

Case 4: Data Science Career Track (Springboard)



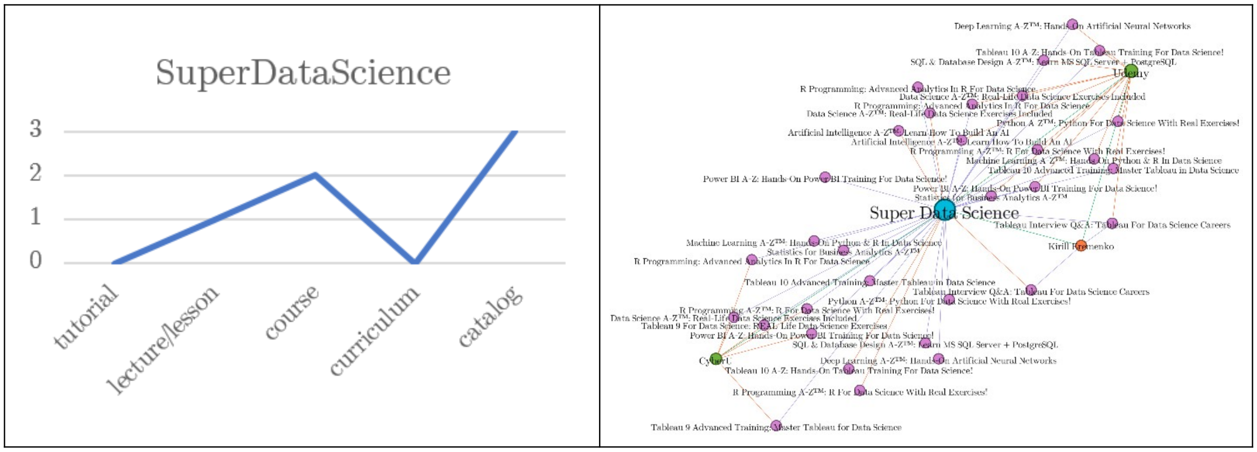
Springboard produces two types of educational offerings: subscription-based, mentor-led "workshops," and free "learning paths." Both are curated from content produced mostly by others besides Springboard themselves, such as from MOOCs, OpenCourseWare, and other resources—much like the OSDSDM. On top of this content, both offering types feature informal tracking on the Springboard website proper, enabling students to mark activities as "complete" at will. The two types differ in that the exact content of the paid workshops is kept hidden until users pay for access, whereas the free learning paths publish their curricula freely; and the workshops provide both a 1:1 mentor relationship and a moderated student community website, while the free option offers neither.

Springboard's liberal reuse of content produced by other actors, without directly collaborating with them in most cases (a notable exception being DataCamp, a self-paced course platform), supports a unique business model, compared to the OSDSM which does not make money from its reused content. Springboard is distinguished even further in terms of its hierarchical rules, in that students can have personalized access to a dedicated mentor. While mentor relationships at Springboard are structured and time-limited, they are notable within the cases reviewed here for directly providing tutorial-level interaction for students. Thus, whereas the program overall capitalizes on an intersection of freely available content, on the one hand, and a self-directed, project-oriented "lifelong learning" activity, on the other, Springboard seems to also provide a unique, more traditional-looking practice in the mentor relationship, supporting a more complex pedagogic discourse.

This relationship has implications for evaluation criteria. In the Data Science Career Track workshop, for example, students are free to "mark complete" whatever they wish and can conclude the program when they choose, so that the mentor relationship provides a check against students' actual progress, while further, a money-back guarantee that students get offered a data science job within six months of graduating from the program—which itself has detailed eligibility requirements—serves as another check, ensuring that students put in a fair level of effort.

Even though the A/R graph shows concentration of focus on curriculum, then, the bundled work happening at the level of “tutorial” is then equally as crucial to Springboard's offering overall, while by merit of its uniqueness, it is also key to assessing power and control relations in the context. After all, Springboard's concept of data science is organized around market demand rather than (and despite liberally using content from) academic research *per se*: the Data Science Career Track workshop is defined by whatever is required to secure a data science job offer, while less career-focused offerings still advise projects be “targeted to a specific client” (Springboard, 2017). The mentor relationship then could hypothetically provide tailored, elaborated transmission of more "powerful" knowledges, leveling access and equalizing power relations. Whether the program does this is not clear from the website or its reviews.

Case 5: SuperDataScience Subscription (SuperDataScience)



The SuperDataScience (SDS) network is densely interconnected, reflecting the fact that this one education vendor resells the same courses in different combinations across three different marketplaces: Udemy, CyberU, and a dedicated SDS website. This case was chosen as representative of the “course marketplace” sub-network identified above, which besides Udemy, CyberU, and SDS, also contains Skillshare and other individuals and individual educational companies.

SDS courses are sold individually, except on the SDS website itself, where a subscription is instead sold for $35/month. Across all sites, prominent advertising links both to other courses within each site as well as back to the primary site. I took this redirecting/funneling behavior to indicate that the context should be properly understood here as the subscription product itself as sold on the SDS website. This means that the catalog of courses must perform as a whole, even while individual courses are exposed to multiple marketplaces and must perform on their own in that sense.

In terms of classification and framing rules, the courses are individually highly structured, but self-paced, with informal “homework assignments” and projects featuring model answers, and, though rarely, simple quizzes. There doesn't appear to be any push toward getting students to post projects to GitHub or similar, as in other cases reviewed above, but rather the evaluatory focus is on qualifying for a certificate of completion at the end and passing personal goals related to comprehension along the way. This is reinforced by course marketing, which lists no prerequisites besides "a passion for success." The content itself is then pitched to practical application in workplace-oriented examples, with a heavy focus on demonstrating tasks in specific software packages, and in turn, the coursework is legitimated on the basis of the practical experience of the lecturer. Above the level of individual courses, there is no program, *per se*, only the catalog, i.e. no ordering is provided. The lecturer however presents himself as friendly and easily accessible via email and comment sections of the various websites; he also hosts an interview-based podcast, provides an email newsletter, and an active social media presence. This rhetorical position helps temper against the impersonal and indeed non-interactive nature of the educational product proper.

Especially when extended by podcasts and articles in the subscription product, the curriculum emphasizes the skills and social positioning of the data scientist, as a technician and job-seeker rather than scientist. For example, over one third of the “Data Science A-Z™: Real-Life Data Science” course is dedicated to data preparation using specific database technologies, while another 10% focuses on presentation skills. Rather, the most technical portion of this course focuses on explaining various forms of regression and ways of performing and assessing them using Tableau and Microsoft Excel. Bernstein's concept of “elaborated code” is then helpful for assessing SDS, as definitely more elaborate in the sense of providing thorough detail about specific technical tasks for example, but not elaborating in the sense of unlocking more abstract, integrating, perhaps theoretical knowledge.

Summary

While the meso-level institutional ecology of U.S.-based higher education demonstrated strong connectivity around an “elite” core of institutions, the micro-level cases add nuance to this view. For example, the OSDSM or Springboard arrangements might be viewed as a network-making "counter-power" to that core network of which they are a part. At the same time, Syracuse University's hybridization approach, while not operating within the network core, might be viewed as being more equalizing or empowering for some students in ways that online education offerings within the core network are not; i.e., in- vs. out-group performance, or discourse, does not follow a fixed pattern. Put another way, pedagogic discourse does not appear to be a function of the status of individual institutions, but of the contexts produced by institutions.

This is precisely where the A/R dimension is intended to contribute, where even if it seems inappropriate to translate a given slope, variance, or other graph metric into pedagogic discourse directly, analysis of A/R still provides a means to visually summarize the nature of differences between contexts, for example the dramatically leveraged focus of OSDSM compared to other offerings. Further, its use enforces a template of questions to be asked across contexts, supporting comparability of findings at that level, even if the graph is not actually drawn. By thus making explicit the arrangement of production and distribution, the economic pressures of each context can be made visible in-context as well as more easily comparable between contexts. As a general conceptual frame, this de-naturalization of the economics of pedagogic context then helps to make visible the ideological assumptions which informed the production of pedagogic discourse. This, finally, bridges analysis of the empirical ground reality of contexts to analysis of pedagogic device.

I next consider how the findings of the study might apply to macro-level analysis of pedagogic device in U.S.-based data science education, and in higher education generally, in the following discussion chapter.

5. Discussion

Introduction

This study began from the premise that studies of higher education have suffered from an analytical gap in recent years related to the mapping of sociology of education onto new institutional forms. In response to this claim, I argue that the study’s findings do point to discourses which, despite having continuity with previous practices, are yet new in socially important and analytically instructive ways, with implications for macro-level analysis of the field of higher education generally as well as for future research.

Macro-Level Implications

Modes of Pedagogic Discourse

Reading the micro-level case findings with the meso-level findings and conceptual framework, it is possible to generalize a typology of *modes* of pedagogic discourse, that is, patterns of discourse which could be expected to reappear in multiple contexts. These are presented in Table 3, where I have titled each with a descriptive/evocative term intended to echo broader terminology and themes in critical literature.

Modes of Pedagogic Discourse	
Virtual University	Echoing the discourse identified in literature review, this mode focuses on extending the traditional mode of production, social relations, humanistic orientation to knowledge, etc. of traditional universities into online modes of delivery. *Example:* Syracuse University/2U, Inc.
Institutional Networking	Directed at crafting political-economic power through the engineering of new institutional arrangements. Uses technology and modes of production strategically to enhance network power. Not directly tied to the state in conventional ways. *Example:* MicroMasters
Entrepreneurial Labor	Educational activity is produced using, and reinforces the regime of, a form of labor that is discursively constructed as self-empowering but which simultaneously conceals economic precariousness—what Neff, et al. (2005) term "entrepreneurial labor." Students take ownership over learning, including in terms of curriculum development and assessment/evaluation, in a socially performative manner. *Example:* Open Source Data Science Master's
Incubator/Guild	Resembles a traditional guild association, or contemporary business incubator, in that it creates social configurations through which "masters" can train "apprentices" in both technical and tacit knowledge, treating data science as craft knowledge. Emphasis on completing "real world" projects as the basis for learning, with job placement and practical experience as the aim and source of legitimation for knowledge. *Example:* Springboard paid workshops; bootcamps
Thought Leadership	Produces educational content which (1) transmits, (2) is legitimated by, and (3) further legitimates, "thought leadership," i.e. the "intellectual evangelis[m]" of popular subject matter experts (Drezner, 2017, 9). Uses less elaborated code, e.g. demonstrates tasks rather than explaining. *Example:* SuperDataScience; Skillshare

Table 3: Modes of pedagogic discourse

Though meaningfully representative of the case, these are not intended to be comprehensive of all discourses even in the field of U.S.-based data science education. Rather, these generalized modes of discourse can be analyzed with respect to the claim made at the start of the study, that new social practices and consequences have been emerging alongside the formal diversity that has come to characterize the field in recent decades. In this respect, the discourses do superficially resemble previously studied concepts such as distance learning (Virtual University), vocational education (Incubator/Guild), or corporate learning (Thought Leadership), but closer inspection of the Entrepreneurial Labor mode, which appears to be the most unique or "new" among the discourses, helps highlight what's new about all the discourses.

The Entrepreneurial Labor mode registers as unique in part because the formal means of its production and distribution, such as networked collaboration technology and open-access content, have only risen to mass use over the last few decades. Equally important, however, is that its performative dimension only registers as "entrepreneurial" in the context of a new economy that has emerged over that same period. The role of economic and especially institutional innovation within the Entrepreneurial Labor mode then points to that same dimension in the other discourses. For example, the Virtual University mode is not only characterized by the addition of distance education to an existing program, but by the intentional and collaboratively-produced blurring of boundaries between distance and on-campus education, as well as between public and private institutional influence—even as the secondary artifacts of the overall activity might in turn become recycled by an Entrepreneurial Labor, Incubator/Guild, or even Thought Leadership discourse involving entirely unrelated actors; or even yet as the actors involved directly in the Virtual University mode simultaneously engage in Institutional Networking discourses in the context of other networks. In this way, the discourses presented above are not only distinct from the previous modes which they resemble in terms of their internal structure, but also and especially in terms of the relations between them. This finding echoes Bernstein's own definition of contexts as about relations "relationships between contexts," as noted in the literature review chapter. In short, the present discursive environment is characterized, if not defined, by an overlapping and interpenetration of discourses, actors, and activities, if not also a multiplicative abundance of the same.

Pedagogic Device

Crucially, while these discourses are produced by different actors, political orientations, economic goals, etc., they do appear to be coordinated by a shared pedagogic device, that is, political-economic or ideological base. Namely, they seem to share a common assumption about a need (or perhaps, in the case of the Institutional Networking mode, opportunity) to somehow scale education, making it more accessible to more students/customers/users, because of shifts in economic reality if not also in the basis or mode of knowledge itself. This first point can be understood in terms of the paradigm of "lifelong learning," a policy concept popularized around the early 2000s which propose that technologically advanced economies require citizens to be formally educated throughout their lives, for the good of nations and themselves. Replying to this policy position, Castells agrees that, practically speaking, education based on physical campuses cannot scale to accommodate such need, meaning that the Internet must be used to meet demand (2009, 4). Bernstein was more pessimistic, claiming that the discourse of lifelong learning signaled a shift toward a "socially empty" pedagogy in which knowledge is rendered as solely economically rather than humanistically valuable, leading to erosion of the autonomy and social meaning of higher education over time (2001, 366).

If this is indeed the ideological orientation of the pedagogic device which organizes the above discourses, then some additional features or phenomena should be able to be hypothesized which future studies could test and explore in more detail. For example, the use of ad-hoc and multidimensional networks of production and distribution, as well as of MOOC- and/or SPOC-style delivery technologies, should be expected to become increasingly routinized and normalized across other subject areas in higher education. Following this, even those contexts which are not explicitly "digital" or "online" should increasingly be *conditioned* by digitally mediated networks, as noted in the Syracuse University case above. This in turn enhances the necessity and urgency of expanding the analytical frame beyond just universities in a contemporary sociology of education.

Policy Context

The policy context of these phenomena sets important stakes for further research. In the last high-level policy report published by the U.S. Department of Education, guidance was provided around the concept of a "'new normal' student [who] may be a 24-year-old returning veteran, a 36-year-old single mother, a part-time student juggling work and college, or the first-generation college student," essentially advocating a lifelong learning discourse of technologically-scaled higher education, justified now as much on the social goal of increasing access to education across social groups as upon stimulating the economic competitiveness of the nation-state (2017, 6). The report goes on to explicitly advocate a de-centering of formal educational institutions in the ecology of public higher education, to be replaced by an ecology in which competency-based frameworks help "create a network of learning that supports students as creators and entrepreneurs, and agents of their own learning" (*ibid.*, 10). The explicit political platform of the current administration, to reduce federal oversight of education while simultaneously allowing federal funds to be directed at a wider range of institutions, serves to increase the momentum of this orientation, even if its political justification for doing so is superficially different.

The contribution of the present study could then be to clarify how the social impact of education can still be assessed and designed in the intentionally disrupted/disrupting institutional ecology anticipated by these policies. For example, whereas efforts to standardize university ratings were undertaken under the previous federal administration, but can no longer apply to the intended new diversity of actors as such, the model of sociology of education presented by this study—particularly its Bernsteinian approach to translating empirical and conceptual research between levels of analysis—should enable a similarly directed instrument for grading and enabling comparison between educational offerings in the new field. This study already demonstrates, for example, that the presence or absence of an elite university does not necessarily ensure transmission of “elaborated” or powerful knowledge, and neither are all the activities of a single institution discursively equal.

Summary

Where the findings of the study meet application to macro-level trends and the policy context of U.S.-based higher education, some insights already appear which point to an institutional ecology for higher education in which universities play an increasing variety of roles, even as they are potentially de-centered as sole authorities over the production of education directly. At the same time, the particularly exploratory, network-oriented, Bernsteinian approach to conducting the study appear promising in their capacity to aid further research into the educational field as it continues to evolve. Indeed, whether guided by federal or state-level government, or else by non-governmental organizations, the educational “marketplace” may increasingly need such reputational information as is made possible by this study’s research approach.

6. Conclusion

In this study, I have attempted multiple distinct aims simultaneously, as laid out by the research problems described in the introduction. I here review these in turn before pointing to further conclusions regarding the significance of the study's results and findings.

First, I described emerging patterns of institutional arrangement in the field of data science education, in order to produce insight into the power and control relations operating in the field, and so respond to a gap in academic literature about data science education. Based on my findings, I provided evidence for an overall displacement of universities as the dominant producers and arbiters of data science education, as well as the simultaneous creation of a new elite core network composed of elite universities, technology companies, and education platforms. In this new “core,” institutional collaboration is high overall, but also characterized by complex recontextualizations for which strong boundaries separate individual collaborative contexts. One practical implication of this is that in terms of assessing any given activity, the role of individual institutions involved is less important than the overall network composition of the context in which the activity occurs. I also provided evidence to suggest that discursively, across contexts and regardless of actors' elite or non-elite status, a disciplining effect of market-oriented educational technology seems to be pervasive, where the particular expression or relative level of influence of this effect merely differs between contexts rather than ever being totally absent. I located this effect in a discourse that has been called “lifelong learning” in the past, but now appears to be recast slightly in current United States higher education policy.

In addition to conducting this directed study of data science education (and indeed, in order to do so), I developed a means by which to apply a specifically Bernsteinian sociology of education lens onto the increasingly networked field of higher education overall. Specifically, I extended Bernstein's conceptual framework of pedagogic “device,” pedagogic “discourse,” and “contexts,” with the network society theory of Manuel Castells, arguing that the multi-level construction of each theory provides a means by which to map concepts between them. I furthermore translated this conceptual framework into a novel methodological approach, combining multiple conceptual levels of analysis with an “agile” graph database to produce a flexible and expressive means for registering, “ontologizing,” and computationally analyzing social data. I argued that the resulting conceptual/methodological framework produced for this study, by capturing the distinctly networked character of the contemporary field of higher education, therefore responded to a general gap in the sociology of education literature related to the theorization of technology, education, and society. It success can be judged by its contribution to producing the uniquely wide and simultaneously detailed and conceptually integrated view of the field of data science education presented above.

The significance of the results from these dual goals of the study are also multiple. First, the study seems to have shown that the field of higher education is now, by nature, generative of discourses in such a way that any study in sociology of education, or of data science education, needs to account in at least a similar, network-oriented manner as was proposed by this study. As shown, Bernstein's conceptual project is uniquely appropriate to such a task, therefore I suggest his work continue to be revisited and updated. In turn, I propose that the agile approach to data management that I applied in this case study would serve future studies well, and is flexible enough to support alternative emphases on e.g. more computationally advanced methods of analysis, given alignment with underlying data model and data collection methods.

Finally, the more general significance of the study might be to point to the possibility of developing, as Bernstein proposed, a "sociology for the transmission of knowledges" (2001, 368). That is, where the focus of sociology of education is implicitly on formal education, Bernstein developed an interest in the more general sociological problem of understanding how the knowledge base of society was developed, maintained, reproduced, etc. Castells has also signaled a sympathy for this question when, in some of his more recent work, he formulated the fundamental basis of power in society to be "the shaping of minds" through "communication power" (Castells, 2013, xix). If it doesn't apply directly to this task, the basic template of Bernstein's project should be taken as a starting point. Certainly, if it is indeed possible to apply the approach developed and tested here to this more general project of investigating a sociology of knowledge, the urgency of doing so is in no short supply.

Bibliography

Ambler, S. W. (2003). *Agile Database Techniques: Effective Strategies for the Agile Software Developer*. Indianapolis, IN: Wiley.

Arsenault, A. H., & Castells, M. (2008). The Structure and Dynamics of Global Multi-Media Business Networks. *International Journal of Communication*, *2*(0), 43.

Barabási, A.-L. (2014). Graph Theory. In *Network Science*. Retrieved from http://barabasi.com/networksciencebook/chapter/2#networks-graphs

Barnett, R. (2013). *Imagining the University*. New York, NY: Routledge.

Baumer, B. (2015). A Data Science Course for Undergraduates: Thinking With Data. *The American Statistician*, *69*(4), 334–342. https://doi.org/10.1080/00031305.2015.1081105

Bennett, S., & Maton, K. (2010). Beyond the ‘digital natives’ debate: Towards a more nuanced understanding of students’ technology experiences. *Journal of Computer Assisted Learning*, *26*(5), 321–331. https://doi.org/10.1111/j.1365-2729.2010.00360.x

Bernstein, B. (2000). *Pedagogy, Symbolic Control, and Identity: Theory, Research, Critique* (Revised Edition). Lanham, MD: Rowman & Littlefield Publishers. Retrieved from https://rowman.com/ISBN/9780847695768/Pedagogy-Symbolic-Control-and-Identity-Revised-Edition

Bernstein, B. (2001). From Pedagogies to Knowledges. In A. Morais, I. Neves, & H. Daniels (Eds.), *Towards a Sociology of Pedagogy: The Contribution of Basil Bernstein to Research* (pp. 363–368). New York, NY: Peter Lang Inc., International Academic Publishers. Retrieved from https://www.peterlang.com/view/product/69296

Bernstein, B. (2009). *The Structuring of Pedagogic Discourse* (Vol. 4). London: Routledge. Retrieved from https://www.routledge.com/Basil-Bernstein-Class-Codes-and-Control/Bernstein/p/book/9780415302869

Cassel, B., & Topi, H. (2015). *Strengthening Data Science Education Through Collaboration [Draft Title]: Report on a Workshop on Data Science Education* (p. 41). Arlington, VA: National Science Foundation. Retrieved from http://www.computingportal.org/sites/default/files/Data%20Science%20Report%20Draft%205-29-2016_0.pdf

Cassel, L. N., Dicheva, D., Dichev, C., Goelman, D., & Posner, M. (2016). Data Science for All: An Introductory Course for Non-Majors; in Flipped Format (Abstract Only). In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education* (pp. 691–691). New York, NY: ACM. https://doi.org/10.1145/2839509.2850558

Castells, M. (2011). A Network Theory of Power. *International Journal of Communication*, *5*(0), 15.

Castells, M. (2013). *Communication Power* (2nd edition). Oxford: Oxford University Press.

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, *69*(1), 21–26.

Corthell, C., & @datasciencemasters. (2017). The Open Source Data Science Masters. Retrieved April 19, 2017, from http://datasciencemasters.org/

Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (4th ed). Thousand Oaks, CA: SAGE Publications.

Czerniewicz, L. (2010). Educational technology – mapping the terrain with Bernstein as cartographer. *Journal of Computer Assisted Learning*, *26*(6), 523–534. https://doi.org/10.1111/j.1365-2729.2010.00359.x

data science - Google Trends. (2017). Retrieved July 11, 2017, from https://trends.google.co.uk/trends/explore?date=2013-07-01 2017-07-01&geo=US&q=data science

Davenport, T. H., & Patil, D. J. (2012, October 1). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*. Retrieved from https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Donoho, D. (2015). 50 years of Data Science (p. 41). Presented at the John W. Tukey 100th Birthday Celebration, Princeton, NJ. Retrieved from http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

Drezner, D. (2017). *The Ideas Industry: How Pessimists, Partisans, and Plutocrats are Transforming the Marketplace of Ideas*. Oxford: Oxford University Press.

edX Inc. (2016, August 15). MicroMasters Programs. Retrieved April 19, 2017, from https://www.edx.org/micromasters

Eybers, S., & Hattingh, M. (2016). Teaching Data Science to Post Graduate Students: A Preliminary Study Using a “F-L-I-P” Class Room Approach (pp. 189–196). Presented at the International Conferences on Internet Technologies & Society (ITS), Education Technologies (ICEduTECH), and Sustainability, Technology and Education (STE), Melbourne, Australia: International Association for the Development of the Information Society. Retrieved from https://eric.ed.gov/?id=ED571590

Finzer, W. (2013). The Data Science Education Dilemma. *Technology Innovations in Statistics Education*, *7*(2). Retrieved from http://escholarship.org/uc/item/7gv0q9dc

Hall, G. (2016). *The Uberfication of the University*. Minneapolis, MN: University of Minnesota Press. Retrieved from https://www.upress.umn.edu/book-division/books/the-uberfication-of-the-university

Hand, D. J. (2015). Statistics and computing: the genesis of data science. *Statistics and Computing*, *25*(4), 705–711. https://doi.org/10.1007/s11222-015-9565-6

Howard, P. N. (2011). *Castells and the Media*. Cambridge; Malden, MA: Polity.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., & Porter, M. A. (2014). Multilayer networks. *Journal of Complex Networks*, *2*(3), 203–271. https://doi.org/10.1093/comnet/cnu016

Little, B., Abbas, A., & Singh, M. (2016). Changing Practices, Changing Values?: A Bernsteinian Analysis of Knowledge Production and Knowledge Exchange in Two UK Universities. In *RE-BECOMING UNIVERSITIES? Higher Education Institutions in Networked Knowledge Societies* (pp. 201–222). Dordrecht: Springer. https://doi.org/10.1007/978-94-017-7369-0

Lucas, H. C. (2016). *Technology and the Disruption of Higher Education*. Hackensack, NJ: World Scientific.

Menard, C., & Shirley, M. M. (Eds.). (2005). *Handbook of New Institutional Economics*. Berlin/ Heidelberg: Springer-Verlag. https://doi.org/10.1007/b106770

Menchik, D. A. (2004). Placing cybereducation in the UK classroom. *British Journal of Sociology of Education*, *25*(2), 193–213. https://doi.org/10.1080/0142569042000205118

Moore, R. (2013). *Basil Bernstein: The thinker and the field*. Routledge.

Morais, A., Neves, I., & Daniels, H. (Eds.). (2001). Video Conference with Basil Bernstein. In *Towards a Sociology of Pedagogy: The Contribution of Basil Bernstein to Research* (pp. 369–383). New York, NY: Peter Lang Inc., International Academic Publishers. Retrieved from https://www.peterlang.com/view/product/69296

Neff, G., Tanweer, A., Fiore-Gartland, B., & Osburn, L. (2017). Critique and Contribute: A Practice-Based Framework for Improving Critical Data Studies and Data Science. *Big Data*, *5*(2), 85–97. https://doi.org/10.1089/big.2016.0050

Pentland, A. (2014). *Social Physics: How Good Ideas Spread—The Lessons from a New Science*. New York, NY: Penguin Press.

Robertson, I. (2007). E-Learning Practices: Exploring the Potential of Pedagogic Space, Activity Theory and the Pedagogic Device. *Learning and Socio-Cultural Theory: Exploring Modern Vygotskian Perspectives International Workshop 2007*, *1*(1). Retrieved from http://ro.uow.edu.au/llrg/vol1/iss1/5

Robertson, I. (2008). Exploring the dynamics that shape teacher's e-learning e-learning practices: An application of Basil Bernstein's pedagogic device. Presented at the Fifth International Basil Bernstein Symposium, Cardiff. Retrieved from https://sites.google.com/site/robboian/Robertson_Ascilite2008_Final.pdf

Robins, K., & Webster, F. (2002). The Virtual University? In *The Virtual University?: Knowledge, Markets, and Management* (pp. 3–19). Oxford; New York, NY: Oxford University Press.

Robson, C., & McCartan, K. (2016). *Real World Research* (Fourth Edition). Hoboken: Wiley.

Selwyn, N. (2010). Looking beyond learning: notes towards the critical study of educational technology. *Journal of Computer Assisted Learning*, *26*(1), 65–73. https://doi.org/10.1111/j.1365-2729.2009.00338.x

Selwyn, N. (2015). Data entry: towards the critical study of digital data and education. *Learning, Media and Technology*, *40*(1), 64–82. https://doi.org/10.1080/17439884.2014.921628

Selwyn, N., & Facer, K. (2014). The sociology of education and digital technology: past, present and future. *Oxford Review of Education*, *40*(4), 482–496. https://doi.org/10.1080/03054985.2014.933005

Song, I.-Y., & Zhu, Y. (2016). Big data and data science: what should we teach? *Expert Systems*, *33*(4), 364–373. https://doi.org/10.1111/exsy.12130

Sowa, J. F. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Pacific Grove, CA: Brooks/Cole.

Sriprakash, A. (2011). The contributions of Bernstein's sociology to education development research. *British Journal of Sociology of Education*, *32*(4), 521–539. https://doi.org/10.1080/01425692.2011.578436

Stanton, J., Palmer, C. L., Blake, C., & Allard, S. (2012). Interdisciplinary data science education. *ACS Symposium Series*, *1110*, 97–113. https://doi.org/10.1021/bk-2012-1110.ch006

Stavrou, S. (2016). Pedagogising the university: on higher education policy implementation and its effects on social relations. *Journal of Education Policy*, *31*(6), 789–804. https://doi.org/10.1080/02680939.2016.1188216

Sung, Y.-K., Lee, Y., & Choi, I.-S. (2016). Contradiction, mediation, and school change: An analysis of the pedagogical practices in the Hyukshin school in South Korea. *KEDI Journal of Educational Policy*, *13*(2), 221–244.

Tang, R., & Sae-Lim, W. (2016). Data science programs in U.S. higher education: An exploratory content analysis of program description, curriculum structure, and course focus. *Education for Information*, *32*(3), 269–290. https://doi.org/10.3233/EFI-160977

Tang, R., & Sae-Lim, W. (2017). Data Science Programs in U.S. Higher Education: An Interview with the Authors. *Journal of EScience Librarianship*, *6*(1). https://doi.org/10.7191/jeslib.2017.1105

Tracy, S. J. (2010). Qualitative Quality: Eight "Big-Tent" Criteria for Excellent Qualitative Research. *Qualitative Inquiry*, *16*(10), 837–851. https://doi.org/10.1177/1077800410383121

Tracy, S. J. (2012). *Qualitative Research Methods: Collecting Evidence, Crafting Analysis, Communicating Impact*. West Sussex: John Wiley & Sons.

Tyler, W. (2001). Crosswired: Hypertext, Critical Theory, and Pedagogic Discourse. In A. Morais, I. Neves, & H. Daniels (Eds.), *Towards a Sociology of Pedagogy: The Contribution of Basil Bernstein to Research* (pp. 339–360). New York, NY: Peter Lang Inc., International Academic Publishers. Retrieved from https://www.peterlang.com/view/product/69296

U.S. Department of Education, Office of Educational Technology. (2017). *Reimagining the Role of Technology in Higher Education: A Supplement to the National Education Technology Plan* (National Education Technology Plan). Washington, D.C.: U.S. Department of Education, Office of Educational Technology.

Varvel, J., Bammerlin, E. J., & Palmer, C. L. (2012). Education for data professionals: A study of current courses and programs (pp. 527–529). Presented at the ACM International Conference Proceeding Series. https://doi.org/10.1145/2132176.2132275

Veaux, R. D. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., … Ye, P. (2017). Curriculum Guidelines for Undergraduate Programs in Data Science. *Annual Review of Statistics and Its Application*, *4*(1), null. https://doi.org/10.1146/annurev-statistics-060116-053930

Yazan, B. (2015). Three Approaches to Case Study Methods in Education: Yin, Merriam, and Stake. *The Qualitative Report*, *20*(2), 134–152.

Yin, R. K. (2003). *Case Study Research: Design and Methods* (Third Edition). Thousand Oaks, CA: Sage Publications.

Appendix: Data Management Technical Reference

Introduction

This appendix documents the database management system in further technical detail, including by defining a data "dictionary" describing each data point tracked, and a schema definition script which can be used to generate the research database structure in a freshly installed instance of OrientDB.

System Description

OrientDB is a multi-modal NoSQL database supporting graph and document-based representations of data. It is open-source software licensed under the Apache 2.0 License (https://www.apache.org/licenses/LICENSE-2.0.html), making it free for commercial as well as non-commercial use. OrientDB can be downloaded from the following URL: http://orientdb.com/download/. The version used for this study was OrientDB 2.2.23 GA Community Edition.

The data model translates an ontology into a formal data structure which can be implemented in the database. This data model was produced in Figure 1: Final data model. OrientDB implements the data model out of "classes," such that a given vertex or edge "instantiates" its class, causing it to inherit that class's properties; and classes can be "parented" by other classes, in turn causing them to inherit the properties of its parent class. The following sections document the definition of this data model further, first as a Data Dictionary describing it semantically, then as a Database Schema describing it technically.

Data Dictionary

This data “dictionary” describes the components of the data model in such a way as to document the detailed properties of each element. The following sub-sections describe the properties of each “class” in the data model. The following three sub-sections describe the three vertex classes defined in OrientDB. Note that the “V superclass” refers to the parent class of Activity and Actor sub-classes.

V superclass

Title	Description
Name	Name of the actor or activity, e.g. "Coursera," "Master of Science in Data Science."
Description	Description provided by the item in question. Could provide the basis for text analysis/classificiaton in future.
URL	URL
DataSource	Code referring to the dataset from which data originates, for example “CC” for Class Central.
Notes	Any additional notes I need to keep that don't fit elsewhere.

Activity

Title	Description
Type	Self-described type. E.g. "MicroMasters" and "Nanodegree" are entered as such, not as "Program."
DegreeLevel	If explicitly defined, enter as e.g. "Master's." Do not infer a value if not provided.
Cost	Cost in USD.
CostType	Per-Activity, Subscription, or NA.
FirstOffered	If known, first year offered.
Delivery	Delivery method. In-Person, Online, Hybrid.
Schedule	Time commitment or scheduling approach. Full Time, Part Time, Self-Paced.
Coordination	Approach to timing. Asynchronous, Synchronous, Self-Paced.
Duration	If fixed, or an average is provided for self-paced content, then store here in the unit provided, e.g. "1.5 months."
AssessmentType	Formal/Informal
OpenContent	Boolean field to describe whether the content created/used by the activity is freely available online. For example, a course might require paid enrollment to offer credit, but its content might be published online via Open Educational Resource licensing.
OpenEnroll	Boolean field to describe whether students must apply and be accepted, or can freely join on their own. Not the same as content being available freely online.
CreditBearing	Boolean field to describe whether formal college credit is offered by the course.
FederalAid	Boolean field to describe whether the course qualifies for federal student aid. Some bootcamps advertise private loans, for example, but these don't count.
DatasetID	Unique identifier from the source dataset, if applicable.

ID	Unique identifier within this dataset.
Class	Text field to describe the class type; helpful for filtering items in Gephi.

Actor

Title	Description
Type	Unlike with activities, a standardized field. Options include: Accreditation Body; College/University; Company—Consulting; Company—Education; Company—Other; Company—Publisher; Company—Technology; Governmental Organization; Non-Governmental Organization; Online Course Platform; Person.
Arrangement	Formal/Informal. Originally designed to support New Institutional Economics frame/concerns, but did not ultimately use.
ForProfit	Boolean field to describe whether the organization is for-profit or not.
Location	If known, an address. Didn't end up filling this out for most items, but it would be interesting in a future study to fill this out further to incorporate a geographic perspective on the field.
ID	Unique identifier within this dataset.
Class	Text field to describe the class type; helpful for filtering items in Gephi.

Database Schema

A database “schema” technically implements the data model within the database. Following is the schema generation script used to generate the research database in its final form. The script is written in a variant of Structured Query Language (SQL) defined by OrientDB. Running this code on a freshly instantiated OrientDB database will reproduce the data model designed for this study.

```
/* ---------------------------------------- */
/* Set database settings for thesis dataset */
/* ---------------------------------------- */
ALTER DATABASE custom useLightweightEdges=false

/* Add properties to vertex superclass for all other vertex classes to
      inherit */
CREATE PROPERTY V.ID IF NOT EXISTS STRING;
CREATE PROPERTY V.Name IF NOT EXISTS STRING;
CREATE PROPERTY V.Description IF NOT EXISTS STRING;
CREATE PROPERTY V.URL IF NOT EXISTS STRING;
CREATE PROPERTY V.DataSource IF NOT EXISTS STRING;
CREATE PROPERTY V.Notes IF NOT EXISTS STRING;
CREATE PROPERTY V.Class IF NOT EXISTS STRING; /* storing explicitly
      helps Gephi/SQL/R downstream */

/* Add property to edge superclass */
CREATE PROPERTY E.Class IF NOT EXISTS STRING; /* storing explicitly
      helps Gephi/SQL/R downstream */
```

```
/* ---------------------------------------- */
/* Create vertices and add their properties */
/* ---------------------------------------- */

/* context class - subclass of V */
CREATE CLASS context EXTENDS V;

/* activity class - subclass of V */
CREATE CLASS activity EXTENDS V;
CREATE PROPERTY activity.Type IF NOT EXISTS STRING;
CREATE PROPERTY activity.DegreeLevel IF NOT EXISTS STRING;
CREATE PROPERTY activity.Cost IF NOT EXISTS STRING; /* USD - OK for
      estimates; out-of-state tuition if in-state is offered */
CREATE PROPERTY activity.CostType IF NOT EXISTS STRING; /* NA, Per-
      Activity, or Subscription */
CREATE PROPERTY activity.FirstOffered IF NOT EXISTS INTEGER; /* Year
      first offered */
CREATE PROPERTY activity.Delivery IF NOT EXISTS STRING;
CREATE PROPERTY activity.Schedule IF NOT EXISTS STRING; /* Self-Paced,
      Scheduled, Mixed. Note that Scheduled + OER:True = self-paced. */
CREATE PROPERTY activity.Duration IF NOT EXISTS STRING; /* Allow custom
      definition here, normalize in analysis */
CREATE PROPERTY activity.Coordination IF NOT EXISTS STRING;
CREATE PROPERTY activity.AssessmentType IF NOT EXISTS STRING;
CREATE PROPERTY activity.OpenContent IF NOT EXISTS BOOLEAN; /* Open
      Educational Resource(s) */
CREATE PROPERTY activity.OpenEnroll IF NOT EXISTS BOOLEAN;
CREATE PROPERTY activity.CreditBearing IF NOT EXISTS BOOLEAN;
CREATE PROPERTY activity.FederalAid IF NOT EXISTS BOOLEAN;
CREATE PROPERTY activity.DatasetID IF NOT EXISTS STRING;

/* actor class - subclass of V */
CREATE CLASS actor EXTENDS V;
CREATE INDEX actor.Name UNIQUE;
CREATE PROPERTY actor.Type IF NOT EXISTS STRING;
CREATE PROPERTY actor.Arrangement IF NOT EXISTS STRING;
CREATE PROPERTY actor.Location IF NOT EXISTS STRING;

/* ---------------------------------------- */
/* Create edges and add their properties    */
/*  - note syntax for defining in/out rules */
/* ---------------------------------------- */

/* ...between any vertex and any other vertex */
CREATE CLASS accredits EXTENDS E;
CREATE PROPERTY accredits.in LINK V;
CREATE PROPERTY accredits.out LINK V;

CREATE CLASS comprises EXTENDS E;
CREATE PROPERTY comprises.in LINK V;
CREATE PROPERTY comprises.out LINK V;

/* ...between any actor and any activity */
CREATE CLASS provides EXTENDS E;
CREATE PROPERTY provides.in LINK actor;
CREATE PROPERTY provides.out LINK activity;
CREATE CLASS develops EXTENDS E;
CREATE PROPERTY develops.in LINK actor;
CREATE PROPERTY develops.out LINK activity;

/* ...between any actor and any vertex */
CREATE CLASS funds EXTENDS E;
CREATE PROPERTY funds.in LINK actor;
CREATE PROPERTY funds.out LINK V;

/* ...between any actor and any actor */
CREATE CLASS collaborates_with EXTENDS E;
CREATE PROPERTY collaborates_with.in LINK actor;
CREATE PROPERTY collaborates_with.out LINK actor;
```