9-8-2022

GROUNDED COMPUTATIONAL ANALYSIS: A HANDS-ON APPROACH TO ANALYSING DIGITAL INNOVATION

Jonas Valbjørn Andersen
IT-University Copenhagen, jova@itu.dk

GROUNDED COMPUTATIONAL ANALYSIS: A HANDS-ON APPROACH TO ANALYSING DIGITAL INNOVATION

Research paper

Jonas Valbjørn Andersen, IT University of Copenhagen, Copenhagen, Denmark, jova@itu.dk

Abstract

As socio-technical processes related to digital innovation are increasingly connected and distributed across geographical, organisational, and temporal boundaries, the methods we use to study them must be adapted to accommodate the greater detail and scope of the phenomenon. Specifically, there is a need to operationalise methods for generating inductive theory of distributed digital innovation from digital trace data. An emerging stream of IS research on computationally intensive inductive theorising lays the groundwork for such methods. This paper builds on this foundation to develop a hands-on approach to operationalising grounded theorizing in computational analysis of digital trace data. The paper first conceptualises trace data of digital innovation as a new research context before articulating an approach to operationalising grounded theory in computational analysis of digital innovation. The application of the grounded computational analysis approach is then briefly illustrated in the context of digital trace data from an online social network before possible directions for further research are laid out.

Keywords: Grounded theory, Computational methods, Digital innovation, Qualitative research

1 Introduction

Empirical research of digital innovation is by nature directed at exploring a novel proposition in the form of the introduction of a new innovation (Bruno Latour, 1991; T. Venturini, 2009), a controversy (Madsen, 2012; Meyer, 2009; Ricci, 2010) or question to be researched such as the introduction of a new technology, actor, idea, or even a shift in the institutional context (Henfridsson & Yoo, 2013). It has recently been established that analysis of digitalization in the broadest sense relate to "a product, process, or business model that is perceived as new, requires some significant changes on the part of adopters, and is embodied in or enabled by IT" (Fichman et al., 2014).

Recent developments in the methodology of computationally intensive qualitative analysis has shown the usefulness of leveraging computational analysis to build inductive theory (Berente et al., 2018; Berente & Seidel, 2014, 2018; Grover et al., 2020), and shown its relevance for analysing digital settings for applying such methods in digital environments (Lindberg et al., 2013; Selander et al., 2010; Vaast & Walsham, 2011). Similarly, research on digital actor networks (Tommaso Venturini, 2012; Tommaso Venturini & Latour, 2010) point to the usefulness of computational methods in unravelling the complexity of distributed innovation processes. This research proposes that the a priori nature of found digital trace data allows researchers to study social interaction processes involving multiple distributed actants at a resolution that is sensitive to individual level characteristics. This allows the researcher to move beyond accounts of a few events of great magnitude to studying the cumulative effects of multiple distributed events of small magnitude (Ruttan, 1954; Usher, 1955). While for natural scientist the availability of large quantities of found data has been commonplace, it represents a great leap forward for the social sciences where "…up to now, access to collective phenomena has always been both incomplete and expensive" (Tommaso Venturini & Latour, 2010). This means it is now possible to develop much more granular methods for analysing distributed socio-technical processes, and specifically digital innovation, in both detail and at scale. Digital traces allow for analysis both in detail *and* at a scope that is commensurable with the unit of analysis of digital innovation. This combination is what sets digital trace analysis apart from other forms of qualitative analysis, where typically a choice must be made between detail and scope.

However, while some initial efforts towards a computational methods for such research have been made, especially within the mapping of social controversies (Bruno Latour, 1991; Okada et al., 2008; Tommaso Venturini, 2012; Tommaso Venturini & Latour, 2010) and computationally intensive qualitative theorising (Berente et al., 2018; Berente & Seidel, 2014, 2018; Grover et al., 2020), the operationalisation of these approaches in researching digital innovation has so far received only sporadic attention in the received literature. Analysing any digitally distributed innovation involves simultaneously observing multiple distributed locations connected through complex and emerging digital infrastructures (Tilson et al., 2010). This in itself presents a challenge for IS researchers, and when adding the need for longitudinal observations of the evolving nature of digital innovations, it is clear that the research techniques traditionally used in physical research settings are increasingly inadequate for analysing digital innovation (Czarniawska, 2004).

The purpose of this paper is to apply grounded theorizing to analysis of digital trace data in the context of digital innovation. The following paragraph will elaborate on the consequences of trace data from digital innovations and explicate this as a new research context. I will then move on to consider the literature on grounded theory building explicating a grounded research practice before describing the methodological underpinnings of grounded computational analysis. The step-by-step operationalisation of the grounded computational analysis approach is then illustrated in the context of a small-scale social network, before moving on to describe its consequences for IS theorizing and briefly outlining possible venues of future research.

2 Trace Data and Digital Innovation as a Research Context

The digitization of once physical environments and practices has been identified as a crucial frontier in researching the organization of social activities in an increasingly digital world (Youngjin Yoo & Lyytinen, 2010). As digital innovations increasingly permeates into physical environments (Y Yoo et al., 2012), an abundance of practices that were once confined to a physical location are now taking place as networked digital innovation. Examples of emerging digital innovation include mobile and digital workplaces, online movements and activism, e-government, distributed product design and innovation settings and open-source communities. What these diverse environments have in common is a surprising inability to answer seemingly simple questions based on existing analytical methods including questions like why do our customers buy our product, how effective are online petitions, and who are our most valuable employees? This presents what one could call a data overload paradox: an explosive increase in the volume and scope of digital trace data leads to an inability to, by means of existing methods, answer seemingly simple questions. The reasons for the data overload paradox are to be found in the materiality of emerging digital innovations. Digital innovations consist of large volumes of digital trace data (Newell & Marabelli, 2015) produced by the increasing digitalization of social contexts (Hedman et al., 2013). Datafication of social actions and relations involves digital agents in the form of algorithms that have recently evolved from processing sequential computational calculations to performing machine learning processes involving interpretation, decision-making and translation. These processes all operate through the medium of digital trace data.

There are at least three defining characteristics to digital traces, which set digital innovations apart from previously known research contexts in the social sciences. First, they are the manifestations of interactions in digital innovations such as status updates, comments, emails, server logs etc. These diverse manifestations of trace data are found data in the sense that they are a by-product of activities rather than produced by a predesigned data collection instrument (Berente et al., 2018; Hedman et al., 2013). Secondly, trace data are relational data as they invariably represent events of actual interactions between socio-technical actors. A final characteristic of trace data is that it is predominantly longitudinal because the events that make it up occur over time (Andersen & Hukal, 2021; Howison et al., 2011).

This means that the volume of available data accumulates at an increasing rate thus reinforcing the process of datafication (Andersen & Hukal, 2021; Lycett, 2013). The increasing datafication leads to an explosion of the scope and range of digital actors as ever more trace data is produced at the same time as increasing global connectivity of information systems and digital infrastructures widens the range of data repositories accessible to researchers. This process leads to an immense increase in data volumes that essentially serves as the fuel that catalyses the activities of digital agents (Andersen et al., 2016).

Consequently emerging digital innovations follow different organizing logics than physical environments (Youngjin Yoo et al., 2010). This has at least two important consequences. Firstly, the programmability and flexibility of the core architecture of digital technology means that digital innovations are continuously shaped and adapted through the social practices they support over time (Henfridsson et al., 2009). This means that research into digital innovation must focus on relations rather than entities and process rather than state. Secondly, people, resources and information are connected in widely distributed and heterogeneous networks that span geographical, organizational and social boundaries and affect multiple social contexts (Lindgren et al., 2008; Y Yoo et al., 2008). For example organizations increasingly rely on external data as previously internal processes are distributed in digital ecosystem environments (Selander et al., 2010). While these consequences of digitization present exciting opportunities, they also present significant challenges to existing research methodology in all phases from data collection and analysis to problems of inference and theory building based on digital traces. Previous studies of digital innovations have emphasized the need for a new methodological approach to studying digital innovations (Bruno Latour, 1991, 1996), but so far attempts have been fragmented and confined to specific contexts (T. Venturini, 2009).

3 Grounded Theory and Computational Analysis

As phenomena associated with information systems are increasingly distributed across geographical, organisational, and temporal boundaries, methods we use to study them are in need of adaptation. The use of computational methods are therefore well aligned with not only the research context of distributed digital innovation, but also relates to recent research commentary proposing that it is of vital importance to IS research to not exclusively observe such digital artefacts from an outside perspective but to provide the perspective of the information system (Grover & Lyytinen, 2015). This means treating IT artefacts as informants that can help identify constructs and their relations and thereby build theory from digital trace data. The method by which theory is constructed in this research is based on a grounded theory approach. It has already been established that grounded theorizing in digitally distributed contexts must involve both be able to describe and classify specific digital innovations and to explain and make predictions that extend beyond a specific research domain (Vaast & Walsham, 2011). It is the ambition of this here to establish a methodology for building such grounded theory using computational techniques.

Grounded theory first gained recognition after the publication of "The Discovery of Grounded Theory" by Barney Glaser and Anselm Strauss (1967). Notably, Bernie Glaser came from a background in quantitative methodology and was trained in qualitative mathematics, a method in which mathematical expressions, such as statistical formulas, can be stated qualitatively (B. G. Glaser, 1998; Strauss & Corbin, 1998). This suggests that bridging the divide between generalizable formal representations and contextualized inductive grounding was an integral component of grounded theory from the very beginning. Since, however, grounded theory has attracted and been the object of many myths about its lack of generalizability including that of the researcher as a 'blank slate' (Urquhart, 2006) and that it leads to the production of low-level theory (Urquhart, 2001).

The following review serves the purpose of mitigating these myths by extending and translating grounded theory methodology to the digital age by proposing a practical framework for building grounded theory using computational techniques. First a general sequence of grounded theory is established by reviewing extant literature on grounded theory. On this foundation a framework for conducting grounded analyses with computational techniques is proposed and finally illustrated with a small-scale empirical example.

3.1 The Process of Grounded Theory Building

Building grounded theory that is based on empirical data is a structured process that can be thought of as involving a number of analytical steps performed in an iterative sequence with the purpose of establishing and refining concepts and their interrelations (Urquhart, 2012; Urquhart et al., 2009). Each iteration aims at identifying and saturating theoretical concepts by sampling, slicing and comparing data, thus adding layers of conceptual abstraction while building on the granularity of previous iterations. This raises the question where to start the process of a grounded analysis. The answer derived from the Glaserian version of grounded theory is remarkably simple: start by becoming an expert on the research domain.

In preparation for collecting the first data, the researcher should immerse into the empirical domain to the point of building general expert knowledge of the phenomenon being studied. This process of domain immersion has been seen as starting from initial 'hunches' (Miles & Huberman, 1984) based on lived experience or more broadly "…sources other than data" (B. Glaser & Strauss, 1967, p. 6). In order to gain such experiential knowledge, the researcher must be deeply familiar with the research domain. Glaser suggests that this initial expert knowledge will guide selection of a 'core category′ of interest within the specific empirical research domain, also referred to as the substantive area (B. G. Glaser, 1978, 1992). This core category will guide collection of the first slices of data (Urquhart et al., 2009).

The initial analytical iteration describes the empirical research domain by establishing narrow concepts and their properties. The first step is to select the area of inquiry guided by expert domain knowledge in the form of a general question or unexplored area within the empirical domain. The first slices of data are then collected within the area of inquiry. These first slices of data are rarely structured a priori, but are broken down into conceptual units with distinct sets of properties through open coding (B. G. Glaser, 1992; Urquhart, 2012). The result is an inventory of narrow concepts describing the area of inquiry.

The second iteration involves interpreting the initial narrow concepts to build substantive theory. Data is sampled using theoretical sampling which represents a key element of grounded theory (B. Glaser & Strauss, 1967). Theoretical sampling involves the successive sampling of data slices based on emergent theory (B. G. Glaser, 1992). To begin with this is based on specific empirical assumptions about the core category, and as more slices or layers of data are sampled, the core category is gradually refined. Process stops when adding more layers of data stops affecting the definition of the core category. This is referred to as 'theoretical saturation' (B. Glaser & Strauss, 1967) and helps determine whether a theory works i.e. whether or not it says something about what is actually going on within area of inquiry (Urquhart, 2012). The resulting substantive theory consists of a set of empirically saturated concepts related to the specific empirical domain in which it is generated (B. G. Glaser, 1978, 1992). Such saturated concepts are generally referred to as a substantive theory because it explains a set of concepts within the specific empirical substantive area (Urquhart, 2001). Presenting substantive theory as the conclusion of grounded research is where some of the criticism of grounded theory for producing low-level theory that is purely descriptive and does not generalize beyond the specific empirical domain in which it was created arise from (Urquhart, 2006).

Therefore, the final iteration aims at building theory meaning 'scaling up' substantive theory to be generalizable over a class of empirical contexts and to be relatable to other theories (Urquhart et al., 2009). Such abstraction is achieved through what Glaser refers to as theoretical coding, which means grouping high-level substantive categories into one or two core categories (B. G. Glaser, 1978, 1992). This increases the density of relations between substantive concepts and adds a layer of abstraction representing a formal, generalizable theory. With every step towards a formal theory, context is necessarily trimmed away from the categories for the sake of transferability and generalizability (B. G. Glaser, 1978).

Figure 1 outlines this process of iteratively building grounded theory from empirical data. Even though for simplicity each iteration is depicted as a singular sequence, in practice each horizontal sequence will be repeated several times until satisfactory conceptualization is achieved. Also, as shown in figure 4, the sequence of grounded data analysis repeats itself in each iteration following a structured set of steps including data sampling, data slicing, data comparison and conceptualization. Data sampling involves mining or collecting a data set based on the criteria developed in the previous iteration. From there the data sample is divided into slices, or layers if you will, depending on the level of abstraction and coded using open, selective or theoretical coding techniques (B. G. Glaser, 1992). Finally, coding is compared between data samples or populations within the sample to reveal conceptual patterns (i.e., concepts and their relations).

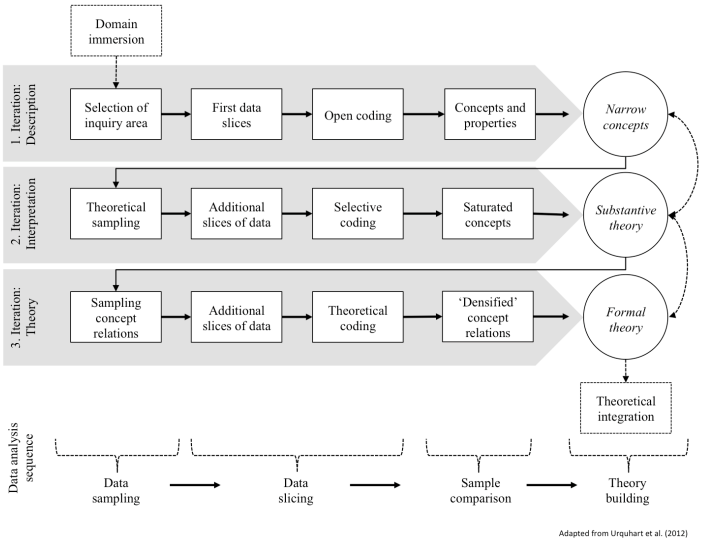

Figure 1. Iterations of grounded analysis

When the data analysis has been conducted, it is of great importance to relate the resulting grounded theory to related theories to ensure that the grounded theory contributes in a valuable way to existing theoretical developments (Urquhart, 2012). This is referred to as theoretical integration and means that the emerging theory should be related to existing high-level theories in the field such as structuration theory (Giddens, 1984; Jones & Karsten, 2008) or actor network theory (Law, 1992; Sayes, 2014). This final step ensures that a direct chain of evidence from each individual slice of data to general high-level theory is established.

3.2 Building Grounded Theory with Computational Analysis

Building such grounded theory using computational techniques, the grounded theory process must be related to existing requirements for digital trace analysis in the social sciences in general (Rogers, 2013) and recent research on mapping digital innovations specifically (Tommaso Venturini, 2012). This section draws the contours of a methodological approach to digital trace data analysis based on the grounded theory process and the materiality of digital traces and environments to propose a research framework for empirical digital trace analysis of digital innovations. Computational data analysis has long been applied in statistics and is also integral to other disciplines in the natural sciences, especially in emerging disciplines such as e.g. genetics and bioinformatics (Jombart, 2008; Kumar et al., 2001). The process of conducting computational data analysis can be described in dive steps: First, the raw observed digital trace data is mined from the research domain. Second, the data is processed in such a way that it fits the analysis. Then, each sample is distinguished by its discrete elements, or variables, providing a 'clean' or 'tidy' data set ready for analysis. Finally, the transformed tidy data set can be processed adding additional variables to each sample (Schutt & O'Neil, 2013). A simplified schematic of the process of a data analysis process associated with the so-called 'non-aqueous fractionation procedure' in molecular biology can be found in Klie (2011). This research shows how the raw data sample, in this case organic material, is transformed through a number of steps adding layers of abstraction and gradually dislodging information from the empirical matter. First, the sample is treated in a way that reveals a particular property of the organic material called the NAF gradient. This 'NAF gradient' is now the data unit being analysed. Then, the gradient material is split into discrete categories through yet another chemical process. Depending on whether or not this process reveals strong enough markers, it is repeated adding new data samples and slices until a saturated set of discrete categories can be measured. Finally, the category measurements are classified into broader groups that are in turn visualized to validate the emerging pattern. The data analysis process applied in these fields include 1) data mining, 2) data unit separation, 3) data discretization and 4) validation and classification and finally 5) conceptualization.

The first step involves mining digital trace data. To be able to investigate as large a variety of diverse questions in as great detail as possible, a high degree of perplexity should be maintained in the raw digital trace data (T. Venturini, 2009). This means the researcher should not simplify the number or nature of patterns to be extracted from the digital trace records. Digital trace data provides a 'found' data source in the sense that it, like organic material, does not need to be constructed with a data collection instrument such as a survey or interview protocol. Instead, trace data can be mined using programming languages such as Python or R to scrape websites or access information systems via data base queries or APIs (Application Programming Interfaces). Therefore, it is important that the data mining process makes as few assumptions about the digital innovation as possible and maintains a highest possible degree of complexity and variety in the data.

The second step involves separating the data into ontological and temporal units. The raw data has such a form that it must be interpreted through a script or application to translate it into a readable format, i.e., in tabular or other structured format. As trace data samples are usually complex and multifacetted, it is necessary to identify and define salient data units. Normally raw digital trace data does not exist in a concise format structured for analysis and is often comprised of data from various sources. This means that data units must be cleaned and separated into units such as e.g., posts, transactions, tweets, updates, profiles etc. In performing this separation, it is important to make sure that the number of voices that participate in the digital innovation and chronological texture is not arbitrarily short-circuited by leaving out salient data units. Therefore, the data sample should be divided into both ontological and temporal units.

The third step refers to discretization of data units into meaningful fractions. This includes evaluating the compatibility of emerging data units by comparing them with already included data units in such a way as to maintain them all in the same setting thus producing a hierarchy or relative positioning of each data unit. This is especially important in tracing the processes by which one emerging ontology redefines or displaces another as is the case in digital innovation (Godoe, 2000; Svahn et al., 2009). In practical terms this means that trace data units should in some way be turned into discrete data to describe the relative strength or position of each data unit.

Having mined, separated, and discretized the digital trace data we can now start identifying and validating patterns from the data. At this point it is useful to employ various data exploration and visualization techniques to support classification of the data (Rogers, 2009, 2013; Tommaso Venturini & Latour, 2010) such as traditional descriptive statistics or other descriptive techniques such as Natural Language Processing (Chang et al., 2013; Landauer et al., 1998) and Social Network Analysis (Granovetter, 1973; Prell, 2012) depending on the stage of theory development. The aim is to validate the discretization and generate pattern classifications for further analysis. As previous research has suggested it can be difficult to draw conclusions from static network analyses (Trier, 2008). Because digital traces are generated through sociotechnical process of interactions with and within digital technology, they are inherently dynamic, why longitudinal analysis should be conducted to empirically validate the emergence of a concepts from trace data.

Finally, once the patterns have been ontologically and temporally validated, the researcher should no longer question their validity as a part of the digital innovation (B Latour, 2004, p. 109). Instead, it is necessary to conceptualize the identified patterns in order to generate and develop insights and theory from the digital trace analysis. This means employing theory and other data sources to develop an explanation for the identified patterns. Figure 2 shows how each computational step corresponds to the grounded analysis sequence discussed in the previous section.

Zooming in for a closer look at the outcome of each data analysis step shown in figure 2, it becomes clear that digital trace data, even though thought of as observational data, undergo a series of computational processes significantly transforming the more or less structured original raw data sample into an analysis ready data set with a very different structure. This process of slicing and moulding the data is underestimated in at least two ways; it represents a significant analytical process which is both time consuming and tedious, and the fact that the original sample is reconstructed for computational analy-

sis means it does not contain direct or 'objective' representations but is prone to a series of biases including being interpreted by the researcher.

In the context of analysing processes of distributed digital innovation, the latter point is especially pertinent as the multi-facetted, distributed, and temporal characteristics of the object of analysis means that extensive domain knowledge is a prerequisite for the researcher to make appropriate analytical decisions (Andersen & Hukal, 2021). This requirement bares remarkable resemblance to the domain immersion described in classic grounded theory literature. Another crucial function of constant domain immersion and re-immersion in grounded computational analysis is the concept of 'iterative conceptualization' (Urquhart et al., 2009) where theory is built in three successive iterations. After each iteration, it is necessary to relate the emerging theoretical constructs to the body of pre-acquired domain knowledge in order to make informed decisions about whether and how to repeat previous steps to increase construct clarity or proceed with the next analytical iteration. That means that analytical reflection and conditioning through domain immersion is effectively the mechanism that distinguishes grounded theory building from variance-based hypothesis testing.

Figure 2 outlines a method for building grounded theory with computational data analysis techniques. The method is formulated as a guide for applying computational techniques in each of the analytical steps, illustrated by solid arrows, for the three iterations of grounded analysis detailed in this section. It shows examples of the reflections required by the digital innovation researcher illustrated by the dashed arrows.

As indicated in the circular conceptualization, the grounded computational method adopts the theory building milestones of a Glaserian grounded theory method discussed in the previous section, narrow concepts, substantive theory, formal theory, and theoretical integration. The way in which each milestone is achieved is where a computational method needs to adopt general data science techniques, specifically explorative data analysis (EDA), statistical models and machine learning algorithms (MLA) and data visualization (DV) techniques (Schutt & O'Neil, 2013). As indicated in figure 5, the first analytical iteration with the purpose of building narrow constructs can be supported through explorative data analysis involving simply describing the characteristics of the data through means of summarization and visualization (Tukey, 1977). Following this, substantive theory consisting of saturated concepts can be built by fitting appropriate statistical or machine learning algorithms such as various clustering, neural network, genetic and deep learning algorithms (Bishop, 2006; Goldberg & Holland, 1988; Marsland, 2014). Relating and visualizing the results of the formal theory building process visualization (Ware, 2012). As illustrated in the models, these computational techniques are just that: techniques. They cannot by any means be automatically applied and need to be treated as any other qualitative coding exercise relying on prior knowledge and analytical conditioning and reflection by the researcher.

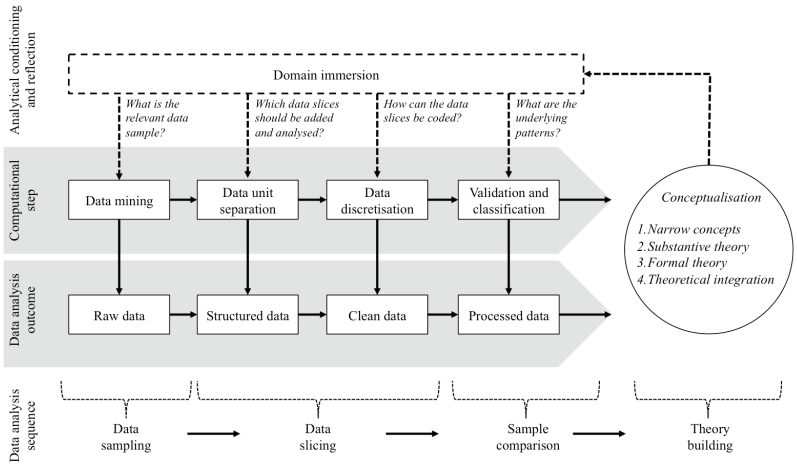

Figure 2. Building grounded theory through computational analysis

The model shown in figure 2 is structured as a single iteration of the grounded theory process discussed earlier. The model should therefore be seen as contingent on the specific research design including the availability of additional, possibly non-digital, data and the patterns discovered in previous analytical iterations. The method outlined in figure 5 can therefore be applied as part of a purely computational study or as a component of a mixed methods research design (Zachariadis et al., 2012) replacing one or more analytical iterations. To exemplify the grounded computational analysis method, the following section describes a brief empirical example of a grounded computational analysis.

4 Empirical Illustration of Grounded Computational Analysis

To provide a thin illustration of the process of how grounded computational analysis can be applied to digital innovation research, the following outlines the key steps in its implementation in a single research iteration aimed to build narrow concepts in the context of a small-scale social network.

A relatively small-scale trace data analysis of a Facebook-like Social Network for PR professionals in Denmark, hereafter referred to as PRnet, was conducted. I had prior to the analysis been thoroughly immersed into both the technological and social dimensions of the research domain, and therefore focused the analysis on the triggers, that make an online social network 'take-off' in the sense that it reaches a critical mass of users and the role of specific users in network growth. The following briefly illustrates the implementation of each step in the grounded computational analysis approach before reviewing key implications and limitations of grounded computational analysis.

Data mining: To analyse the PRnet online social network example I first retrieved a raw digital trace data sample using a SQL database queries. This saved the trace data from the online network into a comma-separated document. The raw and unstructured data represents digital traces generated through interactions on the online social network. The digital trace data represents a seven-month timespan from the formation of the PRnet community consisting of 13,101 connections created by 2,149 members. In the fairly simple PRnet data set, the raw data contained traces of connections between members, member name and affiliation and time intervals for each connection.

Data unit separation: The raw data was then separated into a structured data table where each row represented the formation of a connection on the network using a Python script. Already at this point made some initial decisions are made about the ontological hierarchy inherent in the data by foregrounding connections as the most salient ontological unit. This has direct consequences for the scope of analysis that can be employed at later stages in the process, as it would be possible to choose a different ontological unit such as individual members, organizations etc. Specifically, the raw data was separated into a tabular form with each row representing a connection and each column representing source node, target node, time stamp, and affiliation. This structured format allows for further analysis of our ontological units by separating them into discrete units.

Data discretization: By plotting the number of members with at least one connection at each time interval (1,899 in total) the emergence of the online social network over the first six-month after launch is visualized, thereby identifying the emergence of ontological units, i.e., network members, over time. The left-hand plot in figure 3 summarizes the number of connected users at each indexed time interval. This confirms that indeed take-off in user adoption happens at a specific time interval indicating a "take-off" phase. In fact, by plotting average actor degree in our time series, the phase pattern indicated in figure 6 were reproduced on a continuous scale of node degree. Consulting the data reveals three phases in the emergence of PRnet: the first phase from time index 0 to 17 representing a pre-formation phase, a take-off phase between time index 18 and 40, and a consolidation phase from index 41 to index 100. This way I have and identified at least three discrete units.

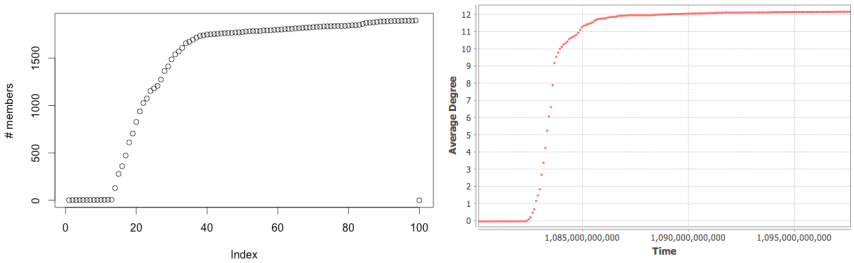

Figure 3. Connected network members and avg. degree over time

However, in order to analyse the mechanisms by which these phases delimited, I decided to measure the degree to which each member is connected to other members. This choice was informed by previous domain knowledge of highly connected central actors in the industry and was chosen as a coding scheme to describe this pattern in the data. The open source social network analysis software Gephi (Bastian et al., 2009) was used to compute connection degree for each member. Degree, also referred to as connectivity, indicates the number of connections for each member. For good measure, both in-degree (number of links directed to each member), out-degree (number of links from each member) and member degree (total number of connections for each member) was computed. The average member degree for the entire network over time is plotted in the right-hand chart shown in figure 3. As the plot shows, average degree increases in a similar pattern to member adoption. This indicates that new members when joining the network connect to existing members rather than form separate disconnected clusters. This confirms that connection degree is a valid metric for coding the emergence of the three phases.

Validation and classification: To validate the degree metric and use it to code members in the network, a graph visualization of the network was generated using Gephi as shown in figure 9. The illustration is drawn making node size dependent on the in-degree to which each member receives connections from other members and node colour is based on outgoing connections. The colour of edges, or lines between nodes, is defined by the in-degree of the target node, i.e., the number of connections pointing to the target member. The graph diagram reveals an emergent pattern where members are located in one of in three spheres depending on their degree of connectivity. I can now distinguish a highly connected core (orange) with an orbiting sphere of medium connectivity (blue) and a green halo of loosely coupled members. Having extracted this pattern from the data and validated it through visualization, a classification of network members emerges. Where to some extent this classification relies on domain knowledge, e.g., the difference in degree is important, the resulting classes and the boundaries between them emerges exclusively from empirical analysis of the digital trace data. In order to further validate the temporal and ontological classifications of the professional network, similar network visualizations for each phase was generated and the average degree computed as illustrated in figure 4.

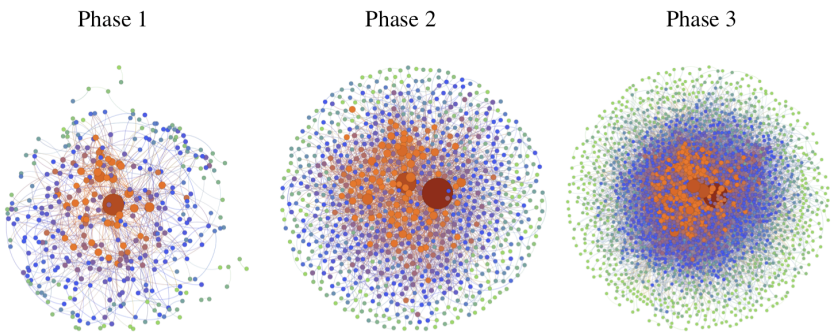

Figure 4. Emergence of the PRnet community

As the network visualizations in figure 4 illustrate, the cumulative position of the network's most central members is reinforced over time as the community grows from the periphery around a central group of highly connected individuals. Interestingly, the material properties of digital trace data, i.e. the temporal separation of relational data units, afford detailed temporal validation of the otherwise static classification of ontological units revealing dynamic patterns in the data. These patterns emerging through grounded computational techniques are the basis for the construction of concepts and their relations and thus ultimately theory building.

Conceptualization: As the analytical iteration illustrated in the PRnet example leads to the construction of narrow concepts using explorative data analysis, conceptualization consists of an inventory of narrow concepts describing the area of analysis. In the PRnet example, this inventory includes three temporal phases and three member categories defined by the connectivity of each member. This initial inventory also hints at some relations between the narrow concepts, specifically at least three patterns, or narrow conceptualizations of the system dynamics, emerge from the data analysis: 1) The PRnet network emerges in three phases including pre-formation, take-off and consolidation phases. 2) Members are divided into three distinct types by being part of either the dense core, connected orbit or the loosely connected halo, depending on the degree to which they are connected to other members. 3) The online network emerged around a highly connected core group and took off with the inclusion of a large group of connected followers.

The analysis does not specifically show whether members transition from the periphery to the center or are a part of a member group by virtue of their profession, seniority, community standing or other attributes. To answer more detailed questions such as which mechanisms drive member adoption and mobility and what type of members are more likely to appear in which tier, more data must be mined and/or sliced based on the empirical conceptualization of narrow concepts, and this data must then in turn be subjected to a similar analytical process.

The PRnet illustration deliberately includes very little text analysis as it aims to illustrate how digital traces can be coded and interpreted as a text. Closer semantic analysis could have been included in the next iteration to saturate the narrow categories derived from this explorative data analysis thus generating substantive theory, just like additional interpretation of outcomes at each step would be expected in the reporting of findings.

5 Scope and Limitations

With regards to the practice and process of conducting grounded research, applying computational methods means adopting a context independent inclusion criterion for coding, in the form of computer code and scripts, that to a certain extent formalizes the link between data sample and analytical coding. This happens in such a way that the coding scheme itself is, beyond the initial data mining, is agnostic of the research domain and specific empirical context if the data format is consistent with the computational coding scripts. The introduction of computational coding scripts has at least two implications for the validity of the emerging theory: First, computational coding scripts allow for visualization and quantification of the link between data samples and theoretical constructs, thus making the chain of evidence more transparent. Also, it potentially leads to transferability of coding schemes across contexts within the scope of the theory, allowing for some measure of empirical reproducibility (Drummond, 2009) and 'theory scoping'.

However, despite these potential benefits with regards to transparency and transferability of grounded theory, computational analysis also introduces several limitations by way of relying on data that are created through existing systems, platforms, and infrastructures. The first limitation stems from the design of the system from which digital traces are mined, and specifically the agendas of the people designing and managing said system. Each digital innovation is built with a set of affordances with the purpose of promoting certain behaviours and deterring others, thus producing digital trace data with an inherent 'design bias'. Secondly, there are at least three reasons why grounded computational analysis is not in itself applicable as a mixed methods research framework (Venkatesh et al., 2013; Zachariadis

et al., 2013). Most prominently, it is not necessarily mixed in the sense that it is usually conducted on a single data source consisting of digital traces and following the grounded approach to theorizing. Also, there are no synergies and interactions between different methods that need to be aligned and exploited in triangulation. Finally, it does not integrate different epistemologies as it follows the grounded theory methods but applies it to digital trace data.

In conducting grounded computational analysis, it is important to reflect on and explicate such design bias. The second limitation originates in the way in which digital trace data is transmitted and recorded in digital data repositories. Being mindful of the materiality of digital traces, they are inherently connected and time stamped. This means that grounded computational analysis is restricted to identifying patterns of processes and relations. As both processes and relations are dynamic and contingent in nature, grounded computational analysis should not be applied to derive universal statements about the properties of certain constructs but explain in detail the contingencies and relational patterns that emerge from the research domain. A final word of caution is that applying a new set of computational techniques to grounded theory building does not mean automating basic analytical reflection or replace basic analytical principles such as the acquisition of domain knowledge, theoretical sampling, and integration to existing theory.

6 Implications for IS Theory Building

Grounded computational analysis provides a method for analysing distributed digital innovation beyond variance-based hypothesis testing. The combination of computational techniques and grounded theory building makes it possible to conduct research aimed specifically at understanding distributed processes related to digitization and the impact of digital technology. The grounded computational analysis approach has at least three implications salient for theorizing distributed digital innovation: a) theory building from the perspective of the digital artefact by accessing digital traces; b) a move from high magnitude to small magnitude theorizing; c) quantifiable qualitative analysis. Each of these implications will now in turn be discussed in greater detail with the purpose of outlining the ways in which a grounded computational analysis approach might influence subsequent theory building.

First, recent commentary calls for research to take the perspective of the IT artefact rather than treating technology as exogenous static entity (Grover & Lyytinen, 2015). In the context of digital innovation this means unravelling the black box of IT artefacts and adopting the point of view of the digital artefact when building theory. The grounded computational analysis approach is an attempt to apply computational tools and techniques in the process of grounded theory building. Grounded computational analysis provides a method for analysing distributed digital innovation beyond variance base hypothesis testing. The combination of computational techniques and grounded theory building makes it possible to conduct research aimed specifically at understanding distributed phenomena related to digitization and the impact of digital technology. This has the added effect of deep immersion into the digital artefact itself thus building theory from the perspective of digital technology, as recently requested in IS research (Grover & Lyytinen, 2015). Secondly, in the context of distributed innovation localized micro-level interactions accumulate into emergent radical transformations through the distribution of access to and control of innovation (Y Yoo et al., 2008).

Following from this, it is important to realize the perils of conceptualizing distributed digital innovation as an act of genius. This leads to theorizing that over emphasizes a relatively small number of actions, which are each of them highly conditional as described previously in this paper. A second danger of adopting a notion of innovation centred around a single protagonist entrepreneur is a notion of change as punctuated and infrequent events of great magnitude (Gersick, 1991; Romanelli & Tushman, 1994). Instead, the conceptualization of distributed digital innovation based on grounded computational analysis builds on a theory of innovation as emerging from changes that "…are numerous, pervasive, and of very small magnitudes" (Usher, 1955, p. 525). This change in analytical scope shifts the focus of conceptualization from entity specific concepts to system centric conceptualization.

This is evident in the empirical example of PRnet, where the initial concepts that are built relate to the general dynamics and properties of the entire social network rather than specific individuals. Deep contextualization and rich description then follows as a way of relating initial narrow concepts in a substantive theory.

Grounded computational analysis favours neither variance research nor adopts a fully interpretivist approach. In this respect, and considering the implications outlined above, grounded computational analysis represents an approach to theory building in information systems that might be labelled as 'quantifiable qualitative analysis'. The analytical approach presented in this paper is potentially useful for researchers wanting to capture the scale and multiplicity of distributed social interactions that constitute digital innovation processes, while maintaining contextualization and semantic texture normally associated with more traditional qualitative research techniques. The only prerequisite is the acquisition by the researcher of basic coding skills. Further research should be conducted to investigate the benefits and trade-offs of automated analytical tools such as those provided by NVivo and Atlas.ti as well as various machine learning techniques for grounded computational analysis.

In summary, grounded computational analysis proposes a method for analysing the emergent phenomenon of distributed digital innovation and unlocking insights relating to the digital artefacts at their core. Consequently, a preliminary set of implications of grounded computational analysis for the IS researcher seeking to implement such an approach might be formulated as follows; 1) acquire domain knowledge 2) take the perspective of the digital artefact 3) question and interrogate information systems as you would documents or human respondents (this includes learning the language of information systems, namely computer coding languages) and finally 4) use computational techniques to build rather than to test theory.

7 Conclusion

The methodological approach presented in this paper has attempted to apply these characteristics to build a novel approach for studying distributed digital phenomena. The grounded computational analysis approach developed in this paper outlines a way forward for combining the connectivity and longitudinal characteristics of such processes in a research approach that is practically applicable to any digital trace record of distributed innovation processes. However, this is merely a first crude step on a long journey to develop new methodological approaches for researching an increasingly connected and digital world. The hope is for this first step to provide some direction for future endeavours in grounded computational analysis of digital innovation.

References

Andersen, J. V., & Hukal, P. (2021). The Rich Facets of Digital Trace Data. In *Cambridge Handbook of Qualitative Research in the Age of Digitalisation*. Cambridge University Press

Andersen, J. V., Lindberg, A., Lindgren, R., & Selander, L. (2016). Algorithmic Agency in Information Systems: Research Opportunities for Data Analytics of Digital Traces. *Proceedings of the 49th Annual Hawaii International Conference on System Sciences*.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *Proceedings of the Third International ICWSM Conference*, 361–362.

Berente, N., & Seidel, S. (2014). Big Data & Inductive Theory Development: Towards Computational Grounded Theory? *Proceedings of the Twentieth Americas Conference on Information Systems*, 1–11.

Berente, N., & Seidel, S. (2018). Don't Pretend to Test Hypotheses: Lexicon and the Legitimization of Empirically-Grounded Computational Theory Development. *Information Systems Research*, 1–36.

Berente, N., Seidel, S., & Safadi, H. (2018). Data-Driven Computationally-Intensive Theory Development *. *Information Systems Research*, *January*.

Bishop, C. M. (2006). Pattern recognition and machine learning. In *Pattern Recognition and Machine Learning*

(Vol. 4, Issue 4). Springer.

Chang, K., Yih, W., & Meek, C. (2013). Multi-Relational Latent Semantic Analysis. *Conference on Empirical Methods in Natural Language Processing*, *October*, 1602–1612.

Czarniawska, B. (2004). On Time, Space, and Action Nets. *Organization*, *11*(6), 773–791.

Drummond, D. C. (2009). Replicability is not reproducibility: Nor is it good science. *Proceedings of the Evaluation Methods for Machine Learning Workshop 26th International Conference for Machine Learning*, *2005*, 1–4.

Fichman, R. G., Dos Santos, B. L., & Zheng, Z. (Eric). (2014). Digital Innovation as a Fundamental and Powerful Concept in the Information Systems Curriculum. *MIS Quarterly*, *38*(2), 329–353.

Gersick, C. (1991). Revolutionary Change Theories: A Multilevel Exploration of the Punctuated Equilibrium Paradigm. *Academy of Management Review*, *16*(1), 10–36.

Giddens, A. (1984). *The Constitution of Society: Outline of the Theory of Structuration*. University of California Press.

Glaser, B. G. (1978). *Theoretical sensitivity: Advances in the methodology of grounded theory*. Sociology Press.

Glaser, B. G. (1992). *Emergence vs forcing: Basics of grounded theory analysis*. Sociology Press.

Glaser, B. G. (1998). *Doing grounded theory: Issues and discussions*. Sociology Press.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory*. Chicago: Aldine Publishing.

Godoe, H. (2000). Innovation regimes, R&D and radical innovations in telecommunications. *Research Policy*, *29*(9), 1033–1046.

Goldberg, D., & Holland, J. (1988). Genetic Algorithms and Machine Learning. *Machine Learning*, *3*, 95–99.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380.

Grover, V., Lindberg, A., Benbasat, I., & Lyytinen, K. (2020). The perils and promises of big data research in information systems. *Journal of the Association for Information Systems*, *21*(2), 268–291.

Grover, V., & Lyytinen, K. (2015). New State of Play in Information Systems Research: The Push to the Edges. *MIS Quarterly*, *39*(2), 271–296.

Hedman, J., Srinivisan, N., & Lindgren, R. (2013). Digital Traces of Information Systems: Sociomateriality Made Researchable. *Thirty Fourth International Conference on Information Systems, Milan 2013*, *38*, 809–830.

Henfridsson, O., & Yoo, Y. (2013). The liminality of trajectory shifts in institutional entrepreneurship. *Organization Science*, *7039*, 1–19.

Henfridsson, O., Yoo, Y., & Svahn, F. (2009). Path creation in digital innovation: A multi-layered dialectics perspective. *Sprouts: Working Papers on Information Systems*, *9*(20).

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity Issues in the Use of Social Network Analysis with Digital Trace Data. *Journal of the Association for Information Systems*, *12*(12), 767–797.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405.

Jones, M., & Karsten, H. (2008). Giddens's Structuration Theory and Information Systems Research. *Mis Quarterly*, *32*(1), 127–157.

Klie, S. (2011). Analysis of the compartmentalized metabolome – a validation of the non-aqueous fractionation technique. *Frontiers in Plant Science*, *2*(September).

Kumar, S., Tamura, K., Jakobsen, I. B., & Nei, M. (2001). MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics*, *17*(12), 1244–1245.

Landauer, T., Foltz, P., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*.

Latour, B. (2004). On using ANT for studying information systems: a (somewhat) Socratic dialogue. In *ANT for studying information systems* (pp. 62–76). Oxford University Press.

Latour, Bruno. (1991). Pour une cartographie des innovations. Le 'graphe socio-technique.' In D. Vinck (Ed.), *Gestion de la recherche, nouveaux problèmes, nouveaux outils* (pp. 419–480). De Boeck.

Latour, Bruno. (1996). Social theory and the study of computerized work sites. *Journal of Technology and Changes in Organizational Work*.

Law, J. (1992). Notes on the theory of the actor-network: Ordering, strategy, and heterogeneity. *Systems Practice*, *5*(4), 379–393.

Lindberg, A., Gaskin, J., Berente, N., Lyytinen, K., & Yoo, Y. (2013). Computational Approaches for Analyzing Latent Social Structures in Open Source Organizing. *Proceedings of the Thirty Fourth International Conference on Information Systems, Milan, Italy*, *1957*, 1–19.

Lindgren, R., Andersson, M., & Henfridsson, O. (2008). Multi-contextuality in boundary-spanning practices. *Information Systems Journal*.

Lycett, M. (2013). "Datafication": Making Sense of (Big) Data in a Complex World. *European Journal of Information Systems*, *22*(4), 381–386.

Madsen, A. K. (2012). Web-Visions as Controversy-Lenses. *Interdisciplinary Science Reviews*, *37*(1), 51–68.

Marsland, S. (2014). *Machine learning: an algorithmic perspective*. CRC press.

Meyer, M. (2009). From 'cold' science to 'hot' research : the texture of controversy. *Papiers de Recherche Du CSI*, *016*.

Miles, M. B., & Huberman, A. M. (1984). Qualitative data analysis: A sourcebook of new methods. In *Qualitative data analysis: a sourcebook of new methods*. Sage publications.

Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification.' *The Journal of Strategic Information Systems*, *24*(1), 3–14.

Okada, A., Shum, S. J. B., & Sherborne, T. (2008). *Knowledge Cartography: software tools and mapping techniques*. Springer.

Prell, C. (2012). *Social Network Analysis - Histori, Theory & Method*. Sage.

Ricci, D. (2010). Seeing what they are saying: Diagrams for socio-technical controversies. *Proceedings of DRS 2010*.

Rogers, R. (2009). New Media & Digital Culture. *Text Prepared for the Inaugural Speech, Chair, University of Amsterdam*, *May*, 1–25.

Rogers, R. (2013). *Digital methods*. MIT Press.

Romanelli, E., & Tushman, M. (1994). Organizational transformation as punctuated equilibrium: An empirical test. *Academy of Management Journal*, *37*(5), 1141–1166.

Ruttan, V. W. (1954). Usher and Schumpeter on invention, innovation, and technologcial change. *The Quarterly Journal of Economics*, *1929*, 596–606.

Sayes, E. (2014). Actor-Network Theory and methodology: Just what does it mean to say that nonhumans have agency? *Social Studies of Science*, *44*(1), 134–149.

Schutt, R., & O'Neil, C. (2013). *Doing data science: Straight talk from the frontline*. " O'Reilly Media, Inc."

Selander, L., Henfridsson, O., & Svahn, F. (2010). Transforming Ecosystem Relationships in Digital Innovation. *ICIS 2010 Proceedings*.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Procedures and techniques for developing grounded theory*. Thousand Oaks, CA: Sage.

Svahn, F., Henfridsson, O., & Yoo, Y. (2009). A threesome dance of agency: Mangling the sociomateriality of technological regimes in digital innovation. *ICIS 2009 Proceedings*. http://aisel.aisnet.org/icis2009/5/

Tilson, D., Lyytinen, K., & Sørensen, C. (2010). Digital infrastructures: The missing IS research agenda. *Information Systems Research*, *21*(4), 748–759.

Trier, M. (2008). Research Note--Towards Dynamic Visualization for Understanding Evolution of Digital

Communication Networks. *Information Systems Research*, *19*(3), 335–350.

Tukey, J. W. (1977). Exploratory Data Analysis. In *Analysis* (Vol. 2, Issue 1999, p. 688).

Urquhart, C. (2001). An Encounter with Grounded Theory. In E. M. Trauth (Ed.), *Qualitative Research in IS: Issues and Trends*. Hershey.

Urquhart, C. (2006). Grounded Theory Method: The Researcher as Blank Slate and Other Myths. *ICIS 2006 Proceedings*.

Urquhart, C. (2012). *Grounded theory for qualitative research: A practical guide*. Sage.

Urquhart, C., Lehmann, H., & Myers, M. D. (2009). Putting the 'theory' back into grounded theory: guidelines for grounded theory studies in information systems. *Information Systems Journal*, *20*(4), 357–381.

Usher, A. P. (1955). Technical Change and Capital Formation. In *Capital Formation and Economic Growth: Vol. I* (Universiti, pp. 521–548). Princeton University Press.

Vaast, E., & Walsham, G. (2011). Grounded Theorizing for Electronically Mediated Social Contexts. *European Journal of Information Systems*, *22*(1), 9–25.

Venkatesh, V., Brown, S. a., & Bala, H. (2013). Bridging the Qualitative-Quantitative Divide: Guidelines for Conducting Mixed Methods Research in Information Systems. *Management Information Systems Quarterly*, *37*(3), 855–879.

Venturini, T. (2009). Diving in magma: how to explore controversies with actor-network theory. *Public Understanding of Science*, *19*(3), 258–273.

Venturini, Tommaso. (2012). Building on faults: How to represent controversies with digital methods. *Public Understanding of Science (Bristol, England)*, *21*(7), 796–812.

Venturini, Tommaso, & Latour, B. (2010). The Social Fabric: Digital Traces and Quali-quantitative Methods. *Proceedings of Future En Seine*.

Ware, C. (2012). *Information visualization: perception for design*. Morgan Kaufmann Pub.

Yoo, Y, Boland, R. J., Lyytinen, K., & Majchrzak, A. (2012). Organizing for Innovation in the Digitized World. *Organization Science*, *23*(5), 13–32.

Yoo, Y, Lyytinen, K., & Boland Jr., R. J. (2008). Distributed Innovation in Classes of Networks. *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, 58–58.

Yoo, Youngjin, Henfridsson, O., & Lyytinen, K. (2010). Research commentary-The new organizing logic of digital innovation: An agenda for information systems research. *Information Systems Research*, *21*(4), 724–735.

Yoo, Youngjin, & Lyytinen, K. (2010). The Next Wave of Digital Innovation: Opportunities and Challenges: A Report on the Research Workshop "Digital Challenges in Innovation Research." *Available at SSRN 1622170*, 1–37.

Zachariadis, M., Scott, S., & Barett, M. (2012). Exploring critical realism as the theoretical foundation of mixed-method research: evidence from the economics of IS innovations. *ICIS 2012 Proceedings*, Article 3.

Zachariadis, M., Scott, S., & Barrett, M. (2013). Bridging the Qualitative-Quantitative Divide: Guidelines for Conducting Mixed Methods Research in Information Systems. *Management Information Systems Quarterly*, *37*(3), 855–879.