This is to certify that the
dissertation entitled

**DIGITALLY-ENABLED ORGANIZATIONAL ROUTINES
AT THE ORGANIZATION-ENVIRONMENT BOUNDARY:
BUFFERING AND THE ROLE OF TECHNOLOGY**

presented by

Derek William Hillison

has been accepted towards fulfillment
of the requirements for the

PhD degree in Business Information Systems

Major Professor's Signature

12/11/09

Date

DATE DUE	DATE DUE	DATE DUE

DIGITALLY-ENABLED ORGANIZATIONAL ROUTINES AT THE ORGANIZATION-ENVIRONMENT BOUNDARY: BUFFERING AND THE ROLE OF TECHNOLOGY

By

Derek William Hillison

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Business Information Systems

2009

ABSTRACT

DIGITALLY-ENABLED ORGANIZATIONAL ROUTINES AT THE ORGANIZATION-ENVIRONMENT BOUNDARY: BUFFERING AND THE ROLE OF TECHNOLOGY

By

Derek William Hillison

Boundary units of an organization uniquely experience the tension between adaptation to environmental variation and maintaining stable outcomes for the rest of the organization. In our world of just-in-time supply chain systems, lot-sizes of one, lean manufacturing and an increasing focus on services, traditional forms of buffering such as queuing and warehousing are not available or less effective. This tension between stability and flexibility must be reconciled in the actions and processes of the boundary unit as reflected in recognizable patterns of action. In addition, the development and organizational adoption of workflow technologies has reduced coordination costs while automating and reifying business rules, both enabling and constraining organizational actions. The assimilation of these workflow systems may fundamentally alter the qualities of flexibility and rigidity in the performance of organizational routines, consequently altering properties of organizational flexibility and adaptation.

I dedicate this dissertation to my wife, who stood by me and supported me through all the snow, dark days, and cold of East Lansing winters as well as the gloriously beautiful summers. She still loves me. I also dedicate this to my Uncle Gary, who always wanted to see me become an author.

ACKNOWLEDGEMENTS

I want to thank Dr. Brian Pentland for his tireless support and patience in advising me through the dissertation process. I don't think I could have done this with a different advisor. I also want to thank my committee for ensuring the quality and impact of my work, while allowing me the time and freedom to explore. Dr. V. Sambamurthy was always willing to guide me to find my story and support it within the literature. Dr. Calantone provided technical, political and social guidance that was very helpful and always appreciated. Dr. Thorvald Haerem provided much of the data collection support in Norway, opened his home to me and even took a valuable Saturday to show me around the more charming places within Oslo. Michigan State University, the Center for Leadership of the Digital Enterprise, The Norwegian School of Management and the Norwegian Research Council provided funding and support for this dissertation. I could not have obtained the data without Thorvald and the help of Compello AS, its president and employees, and Alf Slettemoen. I want to thank my wife, parents, and brother for all the love, dedication and help they provided—it is immeasurable. To all those deserving of thanks who I leave out solely in the interest of saving letters, you have my appreciation and gratitude.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1—Executive Summary

Title: Digitally-Enabled Organizational Routines at the Organization-Environment Boundary: Buffering and the Role of Technology

This research seeks to answer two research questions:

> *Research Question 1: How does a business process at the organization-environment boundary utilize various patterns of action to moderate the impact of environmental variation on process outcomes?*
>
> *Research Question 2: What is the impact of information technology use on the variety found within a digitally-enabled business process?*

These research questions are viewed from a perspective that integrates theory from these three traditions:

- Systems theory and cybernetics advance the ideas of buffering environmental input at the interface between organization and environment and the necessary regulation of variety.
- Organizational routines give a perspective from which to empirically study the emergent patterns of action within a business process.
- The appropriation and assimilation of information systems, focusing on the use of technology within a business process allows the investigation of immediate antecedents and consequences of information technology.

These perspectives are integrated into a general model describing the inputs to a process, the activities that are undertaken within the process, and the outcomes of that process. This model will be evaluated using two methods of analysis on data collected from an invoice processing workflow system.

- Viewing the sequences of action generated within the business process as a Markov model allows familiar statistical techniques based on chi-square tests of homogeneity to evaluate the impact of input variation on the process and its outcomes.
- Alternatively, using sequential variety measures developed from string distance, the input-process-outcome model can be evaluated through regression, path analysis, or structured equation modeling.

Practically, the results of this research increase managerial understanding of the sources and consequences of variety in their processes. Designers of information systems that support business and organizational processes such as workflow, supply chain, or ERP also benefit from a study of the impact of technology on these processes. The methodologies employed in this work can be reapplied in other areas, allowing outcome-based selection and retention of specific characteristics of business process performances. Theoretically, this research enriches at least three traditions of scholarship:

- Evaluation of costs and benefits of Information Technology. By studying the immediate antecedents and consequences of digitally-enabled business process, we gain a better understanding of how the *use* of technology can achieve economic and organizational benefits and costs.
- Consideration of variety in Organizational Routines. While there is an understanding that routines necessarily exhibit variety in their execution, the drivers and consequences of this essential variety are less understood. From some perspectives, process variation is bad (control, audit, TQM), but from others, process variation is good (service quality, responsiveness).

- Extension of sequence methodology. Methodologically, I am applying sequence methods developed in sociology, biology, information theory and social psychology. This research represents both an application and extension of these methodologies, and should result in novel insights and further research in other areas using these techniques.

The main findings of this work are based on data from the processing of 2000 invoices in one organization. In this particular workflow:

- Support for buffering at the boundary has been obtained. Inputs have little relationship to outcomes, but do impact how the process unfolds.
- Automation shows a differential effect on two subprocesses of invoicing. In the data entry phase, it is a substitute for process-based buffering, while in the approval phase, automation is a complement.
- Markov and string distance methods complement each other in that they study the antecedents and outcomes from temporal structures differently.

Chapter 2—Introduction

"...we must accept the coexistence of mutually contradictory phenomenon without trying to resolve the contradiction...new technologies will permit customized manufacture on a mass basis. Rather than being limited by the paradox, they seem to embrace and transcend it" (Davis, 1989).

Overview

Classically, organizational scholars describe mechanisms of insulating and protecting the 'technical core' of an organization from environmental variety and uncertainty by warehousing, demand leveling, and contingent action plans (Thompson, 1967). Boundary-spanning units of an organization experience the challenge of absorbing, managing, and controlling environmental variety (Lynn, 2005; Meznar & Nigh, 1995; Thompson, 1967; Yan & Louis, 1999). The rise of information technology and innovative business models such as just-in-time inventory and custom manufacturing make many traditional methods of insulation less useful.

Actions that are contingently performed based on context or stimulus allow processes to embrace and absorb a given amount of variation, yet remain manageable and achieve controlled outcomes (Ashby, 1956, 1968; Davis, 1989; March & Simon, 1958; Simon, 1996). At the same time, the virtualization of workflow and the advent of inter-organizational information systems have increased the technical structure that boundary-spanning processes must operate within (Basu & Kumar, 2002; Chen, Chen, & Shao, 2003; Georgakopoulos, Hornick, & Sheth, 1995), challenging our understanding of process variety and stability under these conditions.

I develop a worldview that integrates the open system perspective of the organization with the necessary regulation of variety from the environment. I focus upon contingently firing actions and the use of technology within business processes at the boundary of the organization and environment. Viewing these business processes as performative aspects of organizational routines (Feldman & Pentland, 2003) allows the concept of sequential variety to describe the nature of the patterns of action that are expressed (Pentland, 2003a, 2003b). Combining this perspective with cybernetic homeostasis and the use of technology helps describe how a boundary business process can absorb variable inputs from the environment, *and* explore the impact of technology on the process and its outcomes.

The boundary process shown in Figure 1 is a regulator or mediator of variety, represented by the relative sizes of arrows on the input and output sides of the routines. Through the contingent expression of actions, boundary processes obtain a controlled and managed flow of outputs despite the incursion of environmental variety without utilizing traditional forms of buffering such as warehousing, demand leveling, quotas and rationing. Technology and actions become the mechanism through which the buffering of variety occurs.

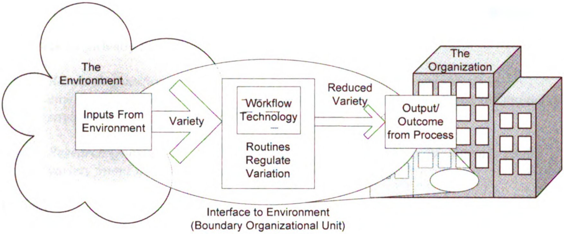


Figure 1: Examining an organizational boundary unit and its routines more closely: It can be seen how the business process absorbs variety through the impact of organizational routines and technology only if we reject the 'black box' perspective of organizational and business processes.

I examine actions performed within an information system poised at the intersection of an organization and its environment. I use process data gleaned from a workflow information system designed for the processing and decision-making surrounding the invoice payment business process. This study gives a rare glimpse into the antecedents and consequences of technology *use* at the business-process level, by isolating the automational aspects of the system and studying their drivers and consequences.

This dissertation represents an innovative study of the antecedents and consequences of sequential variety. Workflow mining techniques give an unprecedented view into the performative aspects of an organizational routine. The workflow system structures and enables the type and sequence of activities that are performed, while providing necessary data for research. This focus on the actual behaviors and use of technology allows the novel use of sequential methods to study the drivers and consequences of temporal action structures. This study represents a unique opportunity to answer empirically the following research questions:

> *Research Question 1: How does a business process at the organization-environment boundary utilize various patterns of action to moderate the impact of environmental variation on process outcomes?*
>
> *Research Question 2: What is the impact of information technology use on the variety found within a digitally-enabled business process?*

Impact

This research is aimed at two main groups of scholars. First, given the rise of interest by organizational scholars in agility, simultaneous exploration and exploration (ambidexterity) and hypercompetitive environments, a reexamination of the classic ideas of buffering and environmental adaptation is warranted. Second, as business processes are increasingly virtualized and digitized, those who study the organizational impact of information technology will have a keen interest in the outcomes of this dissertation and resulting research. Methodologically, this dissertation uses novel methods to measure and analyze sequential processes. Researchers involved in studying the sequential structure of processes such as negotiation, organizational change, or auditing among others benefit from the evaluation and extension of these methods. Table 1 shows these scholarly contributions, along with connections to extant literature.

Contribution	Reference
Focus on a technologically-enabled business process, and its immediate antecedents and consequences	(Kohli & Hoadley, 2006; Mukhopadhyay et al., 1997)
Isolation of the automation aspect of IT use	(Mooney, Gurbaxani, & Kraemer, 1996)
Antecedents and consequences of sequential variety of a work process	(Pentland, 2003a, 2003b)
Application of workflow mining to organizational research questions	(van der Aalst et al., 2003; van der Aalst & Weijters, 2004; van der Aalst, Weijters, & Maruster, 2004)

Table 1: Summary of scholarly contributions

Similarly, there are two main groups of practitioners that benefit from this research: managers and information systems professionals. As managers seek to improve the performance of business processes in the face of environmental change and variety, a better understanding of the antecedents and consequences of process variety may give insight into their attempts to creatively control these processes. Managers, designers and users of workflow technologies need to understand the complex impacts of technology use on the structure and resilience of business processes. For example, real-time workflow analysis systems that are organized around extracting and correlating patterns of action with their antecedents and consequences are core to the development of a 'digital dashboard' for processes (Weske, van der Aalst, & Verbeek, 2004). In addition, managers may need to be reminded of the inherent variety in performing a task in the 'one best way'.

Research Design

The acquisition business process sits at the interface between an organization and its vendors, and contains a subprocess of invoicing (Dunn, Cherrington, & Hollander, 2005). Processing invoices is a perfect setting to study the buffering of environmental variety because of the conflicting pressures of institutional norms and rational management against the flexibility necessary to meet the needs of the vendor and internal constituencies. For example, managers require a process that is controlled and manageable, but variations in the requirements of the vendors and contracts may reduce both consistency and the ability to direct action from above (Baird & Wėisberg, 1982).

I use data extracted from an invoice processing system at a construction company in Norway to evaluate the relationships between inputs, sequential variety, outcomes, and technology. I obtain a log of all the actions and their parameters that take place within the flow of work surrounding the invoice as it is scanned, entered in the system, and approved. The event log is processed into a list of sequentially ordered actions, associated with each invoice. These sequences form the basis of the analysis in this dissertation, allowing the use of multiple complimentary methods to explore the research questions.

Using a multi-method approach increases the amount of effort, but generates a richer, more comprehensive picture of a phenomenon. The two methods I have chosen have traditional uses in their respective areas (Abbott, 1990b, 1995; Sankoff & Kruskal, 1983), have been applied to workflow and business processes (Cook & Wolf, 1998; van der Aalst, 2003; van der Aalst et al., 2003; van der Aalst & Weijters, 2004), yet remain novel in application for the theoretic areas I am targeting. I first view the process as a Markov chain of probabilistically determined actions. Then, I use string distance and multidimensional scaling to understand the sequence of actions within the process and to address the research questions.

The Markov approach views each set of processes as a matrix of transition probabilities between actions (Gottman & Roy, 1990). The sequential structure is determined by evaluating the information that a given action provides about future actions within the sequence (Anderson & Goodman, 1957). Sequences are stratified by the variables relating to inputs, outcomes and the use of automational technology. Log-linear contingency table tests (chi-square) assess the impact of these variables on the temporal structure of the process and the impact of changes in the process on outcomes (Anderson & Goodman, 1957; Bishop, Fienberg, & Holland, 1975; Gottman & Roy, 1990).

For the second approach, the sequences are analyzed by using string-distance techniques (Abbott, 1990b; Pentland, 2003b; Sankoff & Kruskal, 1983) and then scaled for interpretation (Kruskal & Wish, 1978). The sequence is regarded as a string of symbols, and the distance between these strings is computed by counting the number of steps it takes to convert the first sequence into the second. The resulting distances are then scaled using non-metric multidimensional scaling, visualized and then correlated with variables of interest to explore their relationships (Kruskal & Wish, 1978). These scaled distances are used to represent the process in a partial least squares analysis of the relationships between inputs, the process, outcomes, and automation.

Results

Overall, I find support for the input-process-outcome model as buffering the rest of the organization from the variety found in the inputs. Sequences are driven by inputs, and are loosely linked to outcomes. This means that the process is absorbing some of the variance introduced by the inputs. The two methods I use complement each other in that they explore the relationships between antecedents and consequences of the temporal structures expressed in the performance of an organizational routine

The Markov analysis indicates that vendor experience and invoice amount drive heterogeneity in the process. When the sequences are stratified by these variables, the resulting transition matrices indicate that there is a lack of similarity between groups of similar inputs. Using this method, the entry and approval phases of the invoicing routine show differences in their relationship to outcomes. When stratified by the length of time between scanning and full approval, entry sequences are homogenous, but the approval phase sequences were different from each other. The Markov analysis indicated automation as a strong driver of heterogeneity for both the entry and approval phases. For example, none of the transition matrices from four groups of sequences stratified by automation were similar with each other.

While the scaled string-distance approach resulted in dimensions of the process that were difficult to interpret, there was indication that the process was at least partially driven by the input variables. There were marked differences in the patterns of significance and magnitude of coefficients between the entry and approval phases of the invoicing routine, and this was also seen in the partial least squares analysis on the full model. The entry phase utilizes automation as a substitute for buffering through contingent action, while the approval phase uses technology as an adjunct to process buffering.

Outline of Chapters

The remainder of this dissertation is structured into the following chapters: In chapter 3, I review the relevant literature and develop the theoretical model. I then describe the design of this research and the data that I have collected in chapter 4. This is followed by chapters detailing the results of each of the two methods I use : Markov (chapter 5) and string-distance (chapter 6). In chapter 7 I discuss the implications and limitations of the results, and describe future avenues of inquiry.

Chapter 3—Literature Review and Theory Development

Introduction

In this chapter I review the extant literature and describe and support my theory. I begin with a brief orientation to the literature and motivation of my research questions. Then, I discuss the specific relationships between technology use, environment, process, and outcome, supporting these with examples from the literatures of organizational theory and technology impact. I include a reinterpretation of a technology impact study (Mukhopadhyay, Rajiv, & Srinivasan, 1997) to show how the proposed theory can be applied to processes in other organizations such as the United States Postal Service.

Review of Literature

Modern organizations are challenged by conflicting pressures of flexibility and stability as they moderate the flow of resources they use to add value to their outputs. Business processes that exist on the boundary must perform much of this regulation, but many strategies such as JIT inventory and mass customization preclude the use of classic buffering techniques such as demand leveling and queuing. This leaves boundary processes to contingently express different action plans or subroutines in an attempt to reduce the variety present in organizational inputs. Viewing these business processes as patterns of action, the perspective of organizational routines provides a structure to explore the performative aspects of these processes as they respond to variant inputs and moderate the variety to the rest of the organization.

While there have been several examples of empirical research into buffering (Brown & Eisenhardt, 1997; Koberg, 1988), none focused specifically on sequential variety (and performances of an organizational routine) as a measure of contingently driven actions from external stimuli. We know that buffering is a well-accepted feature of successful open systems, and that theory predicts the use of contingent actions as responding to the environment, but there have been few empirical studies (Culnan (1992) as an exemplar) that explicitly study *how* organizations respond to stimuli at the level of the business process. This leads to research question 1:

> *Research Question 1: How does a business process at the organization-environment boundary utilize various patterns of action to moderate the impact of environmental variation on process outcomes?*

Understanding answers to this research question begin with our perspective of the organization. As models of the organization evolved towards open system perspectives, theorists recognized a need to include features of the environment that affect the internal operation and management of the organization (Scott & Davis, 2007). Scott and Davis (2007) define the open systems view of organizations as those that are "capable of self-maintenance on the basis of throughput of resources from the environment".

> "That a system is open means, not simply that it engages in interchanges with the environment, but this interchange is an essential factor underlying the system's viability" (Buckley, 1967), quoted in Scott & Davis (2007)

The first implication of the open system perspective is that there must be a boundary between what we identify as the organization and what exists outside. Second, this boundary acts as a buffer that protects the internal workings of the organization by managing uncertainty and variant inputs from the environment. Third, principles of cybernetic systems can be applied to an open system to give it life and allow the system to learn and react to changes in the environment. Finally, locating the mechanisms of buffering and applying principles of homeostasis at the boundary bring these features of organizational theory down to the level of business processes.

The Organization as an Open System

Figure 2 shows a representation of the open systems view of an organization. At a lower level of analysis, this same figure also can characterize a model of a single business process, as the conversion of inputs into outputs by the interaction of technology and human action (Melão & Pidd, 2000). This is a simplification of the same worldview model presented in the previous chapter.

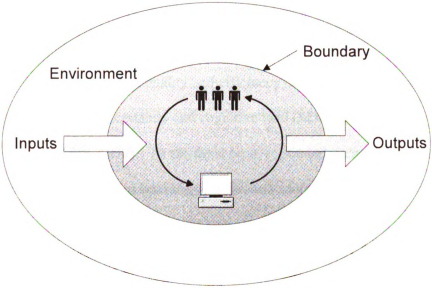


Figure 2: Melao and Pid (2000) show a general model of a business process that also represents the open system view of the organization.

As the organization is opened to the wider environment, managers and workers must deal with the set of events and attributes that is necessarily wider and more varied than exists within the organization. At the same time, norms of rationality drive managers to develop and utilize structures that work towards efficiency and effectiveness in the conversion of inputs to outputs (Spender & Kessler, 1995; Thompson, 1967). This tension complicates the manager's ability to successfully reach expectations of stakeholders in the face of changes in the environment. For example, large variations in demand for technical customer service over the phone make it difficult for managers to achieve target hold times when new products are launched or during wide-spread outages. It is this variation from the environment that creates uncertainty for managers.

Uncertainty creates a question for the organization to answer: How does the need for optimization and control balance against the need for adaptation and flexibility within organizational processes? This question is key to understanding the structures of flexibility and stability in action, and the consequences of enabling and constraining technologies. The conflict arises as uncertainty reduces the ability of the manager to optimize, but the pressures of rationality and efficiency drive decisions and structure towards those that reduce uncertainty and cognitive load (March & Simon, 1958; Thompson, 1967; Weick, 1979). At the same time, continued performance in the face of changing environments requires learning, change, and adaptation (March, 1991).

Boundaries and Buffering

The interaction between the organization and the surrounding environment requires a boundary to organize and identify what is 'inside' and what is 'outside'. As an interface, this boundary is the location of interrelations to "other entities through processes of resource (inputs) acquisition and product/service (output) disposal" (Yan & Louis, 1999). The organizational interface or 'skin', serves not only as a demarcation or identification, but also allows only those inputs that are desired to cross (Simon, 1996), in effect, protecting and buffering the organization from the uncertainty and full variety of the environment. This protection from the environment is what allows homeostatic systems to exist.

Following Thompson (1967), Lynn (2005) defines buffering as "the regulation and/or insulation of organizational processes, functions, or individuals from the effects of environmental uncertainty or scarcity". Koberg (1988) directly studies how two types of organizations buffer their technical aspects of production from environmental uncertainty. Koberg focused only on some of Thompson's types of buffering, relating them to items developed Khandwalla (1974). These were the degree to which units "maintained buffer stocks and reserve supplies of essential material" of spare parts or educational materials (Koberg, 1988).

Koberg (1988) also found that in school settings, decentralization was significantly related to buffering, forecasting, and smoothing, indicating that these techniques were taking place at a lower level than in oil companies, the other organization type she studied. By moving the buffering techniques down to the work unit that can best control uncertainty and requires the most uncertainty management, Koberg suggests schools can succeed despite the lack of technical structure.

Ashby (1956, 1958, 1968) focuses on the cybernetic principles of homeostasis when discussing the behavior of complex systems such as organizations or functional areas within organizations. The law of requisite variety (Ashby, 1956, 1958, 1968) states that the amount of environmental variety that can be dealt with by a system is directly related to the amount of variety in its possible responses. This feature of homeostatic systems applies at multiple levels including the organizational level and the business process level, as an organization can be seen as a nested structure of these processes.

Thompson (1967, p 81) discusses how buffering techniques occur within boundary spanning units but more recently, Yan and Louis (1999) describe how buffering, spanning, and uncertainty management techniques have been pushed down to the work-unit level due to business process reengineering programs, the advent of cross-functional teams, and the introduction of advanced information technologies. To move our understanding beyond warehousing and queuing as buffering mechanisms, we must examine the variety of specific action sequences that a process employs to buffer the organization. A helpful perspective to observe and analyze repeating organizational actions can be found in the concept of the organizational routine. This is considered in the next section.

Patterns of Action

Causal research into organizational outcomes from processes typically has been focused the variance properties of inputs and outputs, treating the generating process as a 'black box' (Melão & Pidd, 2000, 2008; Pentland, 2003b). This happens because the process is often seen as fixed, as it is in many manufacturing contexts. In many cases, this assumption of a fixed process should be challenged, as it precludes learning, adaptation or variability from study. If we reject the 'black box' perspective, we must adopt a view of organizational processes that focuses on the actions that take place, rather than solely their inputs or outcomes. The organizational routines literature provides a perfect perspective for the investigation of organizational actions.

Becker (2004) notes that organizational routines have been characterized as patterns of action. He continues by describing how several authors have defined organizational routine, concentrating on those that embrace a pattern focus. Feldman and Pentland (2003) also discuss organizational routines as patterns of action, and like Winter

(1964) and Koestler (1967), highlight its changeable nature. Table 2 lists these definitions.

Winter (1964) Quoted in (Becker, 2004)	"Pattern of behavior that is followed repeatedly but is subject to change if conditions change"
Koestler (1967) Quoted in (Becker, 2004)	"Flexible patterns offering a variety of alternative choices"
Feldman and Pentland (2003)	"Repetitive, recognizable patterns of interdependent actions, carried out by multiple actors, but they cannot be understood as static, unchanging objects."
Cohen et al. (1996)	"A routine is an executable ***capability*** for repeated performance in some ***context*** that been ***learned*** by an organization in response to ***selective pressures"***

Table 2: Definitions of organizational routines with ostensive and performative aspects of the organizational routine

Feldman and Pentland further to develop the duality of organizational routines: ostensive and performative aspects.

> "Organizational routines consist of two aspects: the ostensive and the performative. The ostensive aspect is the ideal or schematic form of a routine. It is the abstract, generalized idea of the routine, or the routine in principle. The performative aspect of the routine consists of specific actions, by specific people, in specific places and times. It is the routine in practice. Both of these aspects are necessary for an organizational routine to exist" (Feldman & Pentland, 2003).

The ostensive aspects are those understandings of an abstract nature that define the identity of the routine, often including its purpose. The performative aspects of the routine are what happens when the routine is 'executed'. The performances are the sequences of action, performed by various actors and locate them within time and place. In this research, I am focusing exclusively on the performative aspects of the invoicing routine. Even though the performative aspects of a routine are linked to the ostensive aspects and hence can be identified or named with a singular (the invoicing routine), there is an essential variation in the execution routine due to the agency of its participants (Feldman & Pentland, 2003).

This essential variation has been applied to organizational routines (Pentland, 2003a) and was developed from the concept of sequence variety (Abbott, 1990b, 1995; Abbott & Tsay, 2000). Sequential variety is the property of a set of action sequence or performances to exhibit differences in their selection and order of tasks. While this concept holds promise to help scholars understand flexibility in organizational routines and processes, few studies of the antecedents and consequences of sequential variation exist (Pentland, Haerem, & Hillison, 2007).

The tension between variation and stability in business processes is found within organizational routines as well. To reduce coordination costs and allow for consistency in outcomes, routines must be stable. At the same time, routines must allow for contingent adaptation to immediate conditions and deal with participants and tools that may vary in performance and ability. In this way, organizational routines face similar tensions as business processes and organizations themselves in reliability of outcomes and variability in action (Feldman, 2000; Feldman & Pentland, 2003; March, 1991; Nelson & Winter, 1982; Pentland, 2003a; Weick, 1979, 1998). This 'stability in action' perspective represents an important connection between business process management and organizational theory (Singh, Pentland, Yakura, & Hillison, 2009).

The Role of Technology

The impact of IT has often been studied from firm and industry levels, but business-process level studies are less common (Wagner, Beimborn, Franke, & Weitzel, 2006). Focusing on a specific business process and technology allows for a more efficient study of the impact of information technology, as the impact of variance introduced by exogenous or intervening factors is reduced. Kohli and Hoadley (2006) also note the practical value of intermediate or process-level measurement, describing how more detailed measurement allowed firms to better understand the consequences of IT-driven business process reengineering projects. One of the reasons there are few studies may be because of the complex interaction between technology and organizational structure. This complexity is highlighted in the literature as technology is seen as a simultaneous enactment and consequence of organizational structure (Orlikowski, 1992).

Organizational processes become digitally-enabled through the use of communication technologies such as the internet, email, phone, fax, OCR, XML, or coordination technologies such as workflow systems. By virtualizing the processes, these technologies increase the ability of the organization to coordinate temporally and spatially disparate actions within the business process, increasing the possible span and scope of a digitally-enabled organizational routine, while also increasing the transparency and managerial control that is possible over the routine (Overby, 2008). This tension between enablement and constraint of organizational action is another recognition of complexities to be found in the study of technology impact.

The balance of forces driving flexibility, stability, enablement, and control of organizational routines is complicated by the introduction of information technologies. Organizational technologies incorporate implicit models of work that structure future performance and can reify business rules (van der Aalst et al., 2003). Business process reengineering perspectives typically focus on the technological imperative and the enabling aspects of technological adoption and assimilation (Davenport & Short, 1990; Hammer, 1990), while other perspectives focus on the constraints that technologies place on organizational action (Benders, Batenburg, & van der Blonk, 2006; Gosain, 2004). This divergence in the literature makes it difficult to predict the outcomes of technology use, without empirical study. A focus on expressed patterns of action allows the integration of both perspectives in the study of technology use.

Instead of focusing solely on the antecedents and consequences of organizational IT use, this study sees the patterns of action and use of the information system as an intermediate consequence between assimilation and impact. Given the centrality of process variety to theories of organizational routines, learning and buffering, and the complicated and contextualized fit-driven impact of information technologies, it becomes vitally important for researchers to understand the micro-level consequences of IT use. Locating this in a framework of input-process-outcome allows a much finer grained understanding of what IT *use* really is, and allows research to explore consequences and antecedents of use at the same time. This leads to the second research question:

> *Research Question 2: What is the impact of information technology use on the variety found within a digitally-enabled business process?*

Theory and Model Development

In this section, I develop the general world view and theory of input-process-output that guides my research. I begin with one of the archetypical process perspectives given by Melão and Pid (2000), as shown in Figure 2. I then detail the support for my propositions, continuing the literature review and discussing how the model should operate in the context of workflow and invoice processing. I start with the model shown in Figure 3, and move left to right, then top to bottom in my discussion. Beginning with the boundary of the organization and its business processes, I describe how inputs drive contingent actions and the impact of these contingent actions on sequential variety. I then discuss how various levels of sequential variety drive outputs. I conclude with an exploration of how the environment affects technology use, and the impact of technology on process variety and outcomes.

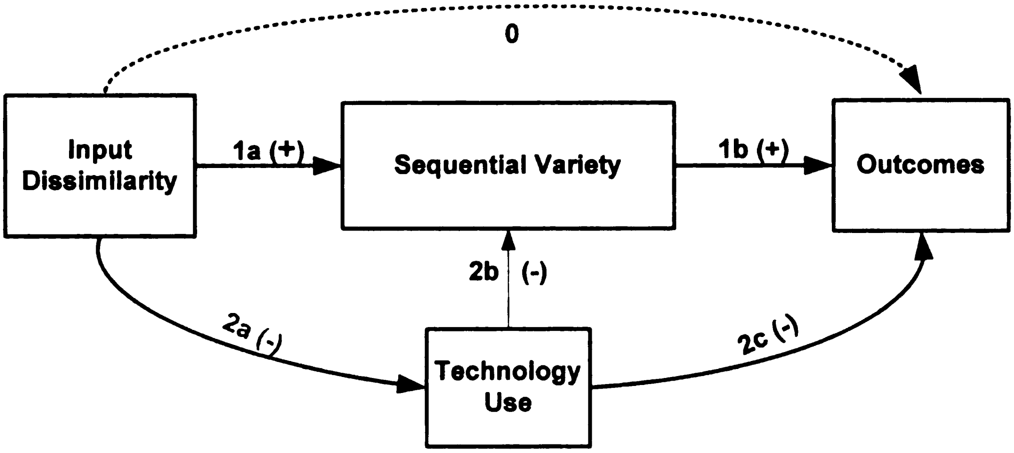


Figure 3: Research Model—the proposed relationships between environmental variation, sequential variety, and technology and outcomes from a business process.

Process Variety at the Boundary

Thompson (1967) lists several forms of buffering the organization from the environment, such as queuing, warehousing, rationing, and demand leveling. While these are appropriate for some manufacturing organizations, they are becoming less available or appropriate for many organizations due to the rise of just-in-time supply chains, custom manufacturing, lot sizes of one, and service provision. Lynn (2005) discusses the ideal relationships between buffering, requisite actions, centralization, and uncertainty. His conceptualization (p. 90) describes traditional forms of buffering as a way for organizational systems to deal with variety in excess of the system's ability to respond through action. In the context of this dissertation, queuing is appropriate for the system to deal with a large influx of invoices at a given time, but demand leveling, rationing, and forecasting seem to be less relevant. Along with Lynn's (2005) decentralized manner of dealing with uncertainty, Yan and Lewis (1999) note that many of the functions previously thought to occur at an organizational level have been pushed downward to the level of the business process by advances in technology and changes in organizational demographics and structure. For the workflow system under study, a perspective that examines the contingent actions that are undertaken seems more appropriate as a regulator of variety than Thompson's (1967) traditional conception of buffering at the organizational level.

March and Simon (1958, p. 45) recognized the inherent flexibility in routines through the contingent use of subroutines, noting that even routines such as those performed on manufacturing assembly lines may have "the character of a strategy rather than a fixed program" (p.45). These subroutines or selected activities would be chosen according to appropriate signals. Since the routines within a boundary unit would be largely driven by stimuli from outside the organization, we should see variation in these routines based on changes in these signals from the environment. The performances of these routines, enacted in the boundary organizational unit, should achieve a variety equal to those features in the environment important enough to warrant differences in execution of the routine (Ashby, 1956, 1958, 1968). In this way the environment of a business process or routine may introduce variation in the actions or their order within the process, leading to the following proposition.

> *P1a: Higher levels of environmental variation will be associated with higher levels of sequential variation within a business process.*

The Relationship between Process Variation and Outcomes

Theory that predicts the consequences of process variation also may indicate a complex relationship with outcomes, given the diversity of ways variation is viewed in literature (Pentland, 2003a, 2003b). It can be a harbinger of lower quality in manufacturing contexts (Oakland, 1999) or an indicator of higher quality in service provision (Leidner, 1993). The nature and results of this study gives some capability to resolve these contradictions.

The processing of invoices may be more similar to services than to manufacturing contexts, so variation in this process should be positively associated with outcomes related to qualitative measures such as success, failure, or meeting a deadline. At the same time, the institutional and legal norms such as accounting standards and internal controls would tend to make extremely variant processes more costly (Dunn et al., 2005). The balance between control and efficiency of a business process often determines its performance, reliability, and cost.

There is a distinct difference here between predicted systemic and performance-level effects. At the systemic level, more process variety indicates a greater ability for the routine to deal with environmental variation. At the level of the individual performance, I predict qualitative measures of outcome to be higher with more sequential variety. These qualitative and systemic benefits from increased sequential variety might come at a quantitative expense of increased processing time or cost. This is an interesting implication of differing directions of effect based on the level of analysis. This leads to a second proposition:

P1b: Higher levels of sequential variety will be weakly associated with a longer completion time when measured at the level of individual sequences.

Technology Enablement and Constraint

At the operational level of business process, technologies typically are given, often beyond the discretion of the individual actor, and yet there must be enough flexibility to meet the needs of that particular business process or routine. An information system has inputs that can be handled automatically because of their formatting and qualities, but there may be inputs that fall outside of this specification that call for special handling as exceptions. The quality and characteristics of inputs can impact an organization's ability to use technology in a business process (Mukhopadhyay et al., 1997). Automated mail sorting technology can read the addresses on a variety of machine written and bar-coded mail, but has difficulty reading hand-written addresses and mail damaged by water. These must be hand-sorted because of the joint characteristics of the input and the technology.

Culnan (1992) found a similar situation with mail handling at the U.S. Senate. Form letters were devised for specific issues and constituencies, and it took many letters regarding new issues to generate a new response format. The Senate used several information systems to generate the form letters and also for other correspondence. After the incoming mail was categorized, a combination of information systems and organizational work was completed to respond to or ignore the various mail inputs.

In Culnan's (1992) case, inputs that were common and recognized were able to be processed by existing form letters within the system. If the letters were outside of this specification, they were ignored, collected for future use, or handled as an exception. The number of responses matched the different types of inputs that the joint features of the organization and technology allowed within the correspondence function. Senators were protected from too much mail by a system that categorized similar inputs, responded appropriately, and learned to add new responses as needed.

One implication of the cybernetic view of systems and information theory also can bring understanding to this issue. The law of requisite variety requires variety in the responses equal to that found in the environment of the system (Ashby, 1956, 1958, 1968). The information content found within the set of available responses is a function of how many responses there are and how many types of signals from the environment are needed to determine the correct response. Automational technologies, such as those found in workflow software, use these signals to fire without decision-making by humans.

To conserve human attention (Simon, 1973), the information system should automate those tasks where the information needs for decision-making can be determined in advance and match exactly the needed systemic response. This implies that those tasks that should be automated are those that do not require additional information or decision-making to direct the correct task. This set of actions must, by definition, be lower than is possible in a human-automation hybrid system at a steady state, given that the amount of variation in the environment and inputs is wider than can be predicted due to limits of human prediction and cognition. This suggests another proposition:

> *P2a: Environmental variation decreases the use of technology for a given business process.*

As noted earlier in this chapter, the relationship of technology to sequential variety is complex. Technology both enacts and is shaped by organizational structure (Orlikowski, 1992). The classic tension between organizational flexibility and stability is complicated by the introduction of technologies that may fundamentally constrain and enable organizational action. This is in addition to the impact this tension has on organizational structure.

This complexity is also present in the literature, as different groups of scholars highlight the equivocal nature of technology in how it impacts on organizational processes. For example, Poole and Desanctis (1990) describe how an information system can be subverted for purposes not in its *spirit* by users while Gosain (2004) highlights the isomorphic pressures from the controlling aspects of an information system.

Business process reengineering perspectives typically focus on the technological imperative (Orlikowski, 1995) and the enabling aspects of technological adoption and assimilation (Davenport & Short, 1990; Hammer, 1990; Lee & Dale, 1998; O'Neill & Sohal, 1999), while other perspectives focus on the constraints that technologies enact on organizational action (Benders et al., 2006; Gosain, 2004). Typically, each perspective's core focus is on either enablement or control (Singh et al., 2009). This complicates our ability to predict, and requires some additional context to understand how these effects will impact this study. The design of features within a specific class of information system can facilitate the development of a contextually relevant prediction. I discuss these next.

Business processes that utilize workflow systems experience more structuring of sequence and action because an implicit model of the process becomes the basis of work design (Georgakopoulos et al., 1995; van der Aalst et al., 2003). While recent advances have increased the flexibility in the handling of exceptional cases and variable processes (Carlsen, 1997; Narendra, 2004), not all workflow systems are designed this way. In fact, managers may wish to specify the actual workflow pattern explicitly to exert control over the process. Business rules may be implemented in the system by automating tasks based on predetermined criteria. For example, an invoice may be automatically approved if it is under a certain dollar amount, or invoices for specified vendors may be passed directly to the required approver. These indicate that the use of technology would be reflected in the patterns of action that are expressed, specifically as a reduction of process variance. Thus, the following proposition:

> *P2b: Increased use of technology will be associated with lower sequential variety for a given business process.*

Brynjolfsson and Hitt (2000) implicitly recognize that it is the structures of use that generate value, because achieving benefit from information technology requires complimentary organizational action. Soh and Markus (1995) outline a process model of IT business value creation, synthesized from integrating several extant research models. Three overall information technology processes are distilled: IT conversion, IT use, and competition (p. 37). The use process focuses on the connection between IT assets and IT impacts, contingent on how the IT is used appropriately in context.

From the perspective of the business process, Mooney et al. (1996) describe three main dimensions of IT business value, focusing on automational, informational, and transformation aspects of the technology in use. Workflow systems may be installed as part of a business process reengineering transformation, but once in place the effects are mainly from the automation and informational features of the technology. This research focuses primarily on the impact of automational features of a technology.

Studies have linked information system use with organizationally important outcomes on revenue and mortality in a healthcare setting (Devaraj & Kohli, 2003) when studied at the level of the business process. Mukhopadhyay, Rajiv, and Srinivasan (1997) also study the impact of IT use on a business process, finding that IT use increased the throughput of mail, with appropriate inputs. All of these studies point to increased value from IT use, contingent on context and appropriate inputs.

> *P2c: Technology use will be associated with lower cost, shorter completion times for a given business process.*

Application to Mail Handling

An empirical example of the theory I use can be found through a reinterpretation of Mukhopadhyay et al. (1997). They examined the impact of a new mail reading and sorting technology at a United States Post Office. This technology increased the speed of sorting and reduced errors in almost all cases except mail that was wet or poorly hand-addressed. The technology was unable to machine-read the address on mail with these properties, and they had to be sorted by hand.

As a system, a different response was developed for each of the types of variant inputs, as predicted by the law of requisite variety. Extending the example given by Mukhopadhyay et al., we can gain insight through the impact of later events on the postal system in the United States. In 2001, someone mailed anthrax spores to many people and some became infected and died. The impact on the mail system was immediate and widespread, as people were afraid to open mail; some mail sorting services were halted while decontamination was completed by people in biohazard suits. In response, the US mail service implemented testing for anthrax in mail at all processing centers (Klatell, 2006). Again, in terms of inputs, the anthrax contaminated mail was variant beyond the system's ability to handle them, and a new set of actions was implemented to allow the system to operate safely to detect and quarantine contaminated mail.

This scenario demonstrates the theory developed in this chapter. It is an example of a system where the inputs drive the changes in process, resulting in improved outcomes. In addition, the role of technology as one of the factors that both enables and constrains actions is clear. In the next chapter, I describe the research design that will validate a contextualized model derived from the theory seen in Figure 3.

Chapter 4—Methodology and Research Design

Introduction

This chapter outlines the collection and processing of data, connecting the theoretic model in the previous chapter to the context of invoice processing and to the methods I use to evaluate the research questions (Figure 4). I seek to understand how signals from the environment (inputs) drive heterogeneity in processes, and how this affects outcomes from the process. Also, I want to discover how automation affects the heterogeneity of the process and the impact of information technology use.

I conclude this chapter with an overview of each analysis performed in subsequent chapters. One important methodological contribution of this work is the development and evaluation of different measures of sequential variety in organizational processes. Table 3 lists these in the order of their increasing inclusion of sequential information. These have been utilized in a stream of related work with similar data, but from a larger set of four organizations (Pentland, Haerem, & Hillison, 2009a, 2009b, 2009c, 2009d).

Research Design

The research design calls for comparing variation in several sets of variables: those related to environmental factors, the sequence of actions in the routine, the amount of automation, and measures of outcome. The environment must be suitably defined and measured; the routine must be situated in an organizational unit that acts as an interface between the environment and the rest of the organization; and the outcome variables must be measurable and linked to the specific behaviors found in the routine. To infer causation, temporal precedence must be established, some form of mathematic relationship must be found, and this must be supported by a plausible story explaining the relationships between the variables (Bollen, 1990; J. Cohen, Cohen, West, & Aiken, 2003; Kenny, 1979)

To explore the relationships between environment, process, outcomes, and technology, I begin by using conditional probabilities of the events in the sequences, treating the organizational routine as a Markov process that varies in response to contextual variables (Gottman & Roy, 1990). This views the variety in the process as the set of possible states and the transitions between these states (Ashby, 1976). For the second analysis, I use optimal string matching techniques to calculate a distance of each sequence from each of the others to create a distance matrix (Sankoff & Kruskal, 1983), measuring the amount of sequential variety by analyzing each sequence of actions in relation to the others (Pentland, 2003a, 2003b). Variety in this analysis seems closer to a measure of dispersion around a set of modal sequences, and may be thought of as the 'standard deviation' of a set of sequences in some ways. As noted above, Table 3 shows these methods and others in terms of increasing amounts of sequential information used.

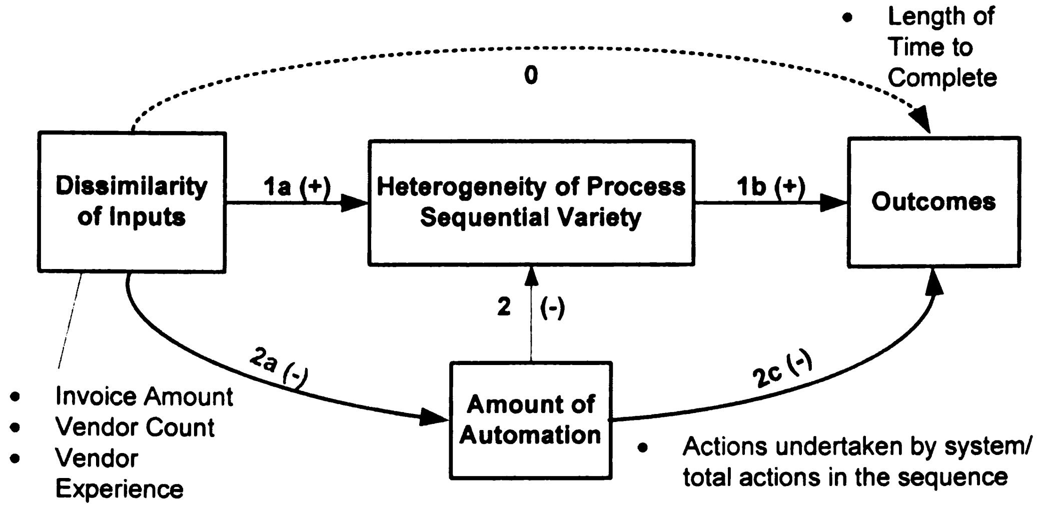

• Length of Time to Complete
0
Dissimilarity of Inputs
1a (+)
Heterogeneity of Process Sequential Variety
1b (+)
Outcomes
2 (-)
2a (-)
2c (-)
Amount of Automation
• Invoice Amount
• Vendor Count
• Vendor Experience
• Actions undertaken by system/ total actions in the sequence

Figure 4: Analysis model, showing contextualized variables and relationships

		Strengths	Weaknesses	Citations
Increasing temporal information→	Lexicon Size	Easy to obtain	Very coarse measure of variety	(Pentland, 2003a, 2003b)
	Lexicon Distribution	Can discriminate between processes	No sequence information used	(Pentland, 2003a, 2003b)
	Entropy	Uses relative probability of execution	No sequence information used	(Ashby, 1956, 1958, 1968; Shannon & Weaver, 1949)
	First Order Markov	Well established, used in variety of fields	Omits information about longer sequences	(Ashby, 1976; Gottman & Roy, 1990; Pentland, 2003a, 2003b)
	Higher Order Markov	Uses additional information about longer sequences	Math gets difficult...harder to prove better fit than lower order	(Gottman & Roy, 1990)
	String Distance	Well established in a variety of fields Uses the entire sequence of actions	May be less appropriate for study of organizational processes Conceptually, what do insertions, deletions and substitutions mean for organizational processes?	(Abbott, 1990b; Abbott & Hrycak, 1990; Abbott & Tsay, 2000; Pentland, 2003a, 2003b; Sankoff & Kruskal, 1983)

Table 3: Methods for measuring sequential variety

Research Site and Data Collection

The invoice processing routine lies at the interface between vendors and an organization, making it a perfect opportunity to evaluate the antecedents and consequences of process variance. Despite institutional and legal norms regarding the form and general process that must be followed, there is a large amount of variability in how organizations can make and document the decisions regarding payment of the invoice (Dunn et al., 2005).

This research uses data collected from an invoice processing workflow system in use at a construction company in Norway. Invoices most commonly enter the system on paper or less often via an electronic portal. If the invoice is on paper, it is scanned and optical character recognition is performed for initial data entry, while some information must be entered manually. Invoices can be immediately sent to the financial system for payment; others require multiple approvals, thus the number of approvals can be larger than the number of invoices. Once the approvals are complete, the invoice is paid. An overview of the invoice process as conceived by designers of the workflow system is shown in Figure 5 (Compello Software, 2007). The data obtained from the invoice processing system is not immediately ready for analysis, as a significant amount of extraction, conversion and transformation is needed to obtain useful data.

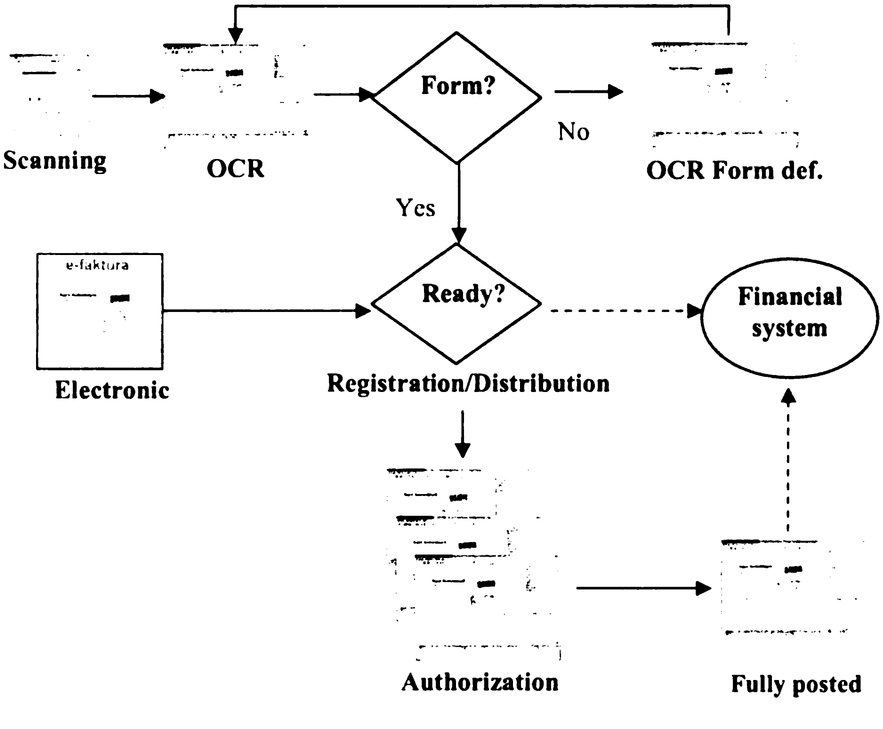


Figure 5: Flowchart of the invoice processing routine

Workflow mining techniques are utilized on the action logs that software provides (Agrawal, Gunopulos, & Leymann, 1998; Agrawal & Srikant, 1995; van der Aalst et al., 2003; van der Aalst & Weijters, 2004; van der Aalst, Weijters, & Maruster, 2004). The resulting data provide the input variables, sequential information, amount of automation, and outcome variables for both the string-distance and Markov analyses. To obtain the sequences, the event log was parsed, and each entry and approval sequence was extracted and linked to the related invoice. An example event log is shown in Table 4, and processed sequences are shown in Table 5.

Invoice #	Phase	Action	Code
202132	Entry	Enter invoice no.	8
202132	Entry	Enter invoice date	7
202132	Entry	Enter due date	6
202132	Entry	Enter amount	3
202132	Entry	Enter currency	4
202132	Entry	Enter document type	5
202132	Entry	Enter vendor account	20
202132	Entry	Enter period	10
202132	Entry	Enter text	12
202132	Entry	Approve	1
202132	Approval	Enter period	10
202132	Approval	Enter currency	4
202132	Approval	Enter amount	3
202132	Approval	Enter text	12
202132	Approval	Distribute to approver	23
202132	Approval	Distribute to approver	23
202132	Approval	Notify	25
202132	Approval	Enter account	2
202132	Approval	Enter Tax-code	11
202132	Approval	Enter value dim. 1 (DP)	13
202132	Approval	Approve	1
202132	Approval	Approve	1

Table 4: Example event log from workflow system

Identifier	Phase	Action Sequence
112568	Entry	8, 7, 6, 3, 4, 5, 20, 10, 12, 1, 9
112569	Entry	8, 7, 6, 3, 4, 5, 20, 10, 12, 1, 9
112573	Entry	8, 7, 6, 3, 4, 5, 20, 10, 12, 1
112568	Approval	10, 4, 3, 12, 23, 23, 25, 2, 11, 28, 16, 1, 13, 23, 27, 25, 15, 1
112572	Approval	10, 4, 13, 3, 12, 14, 23, 23, 25, 2, 11, 23, 16, 25, 1, 1
112573	Approval	10, 4, 3, 12, 13, 14, 23, 23, 23, 23, 25, 2, 11, 15, 25, 1, 1, 23, 16, 1, 1

Table 5: Sequences extracted from the workflow event log

I have data from the system's first installation in 2001 through July 2007, resulting in 58,000 invoice sequences that are available for analysis. A random sample of 2000 invoices from June 2005 through May 2006 was selected for further analysis in this dissertation. I chose this time period to for two reasons: an attempt to avoid start-up learning effects and to center the data on year end (December 31). Because the data source was a construction company in Northern Europe, I expected strong seasonal effects and wanted to minimize and control their impact while still allowing a large sample. Centering the sample on year end allowed both the split-half and four-group tests in the Markov analysis to help isolate seasonal effects from the systemic variation I was seeking. The number of invoices was set at 2000 because this was close to the technical limitations of string-distance analysis.

Evaluation of the theoretical model presented in chapter 3 requires that the concepts and relationships must be contextualized and operationalized, as presented in Figure 4. Table 6 defines and outlines the variables for environment, technology use, and outcomes specifically for the invoice process and research site. I discuss their operationalization (Table 7) in the next section. Since I am interested in how 'different' each invoice process is from the others, scaling, clustering, or stratification techniques are used on the set of environmental and outcome variables. Because I am utilizing multiple analysis methods, I include heterogeneity of the process (Markov approach) and sequential variety (string distance approach) as measures of the process itself.

Variable Type	Definition
Environment	How many times has this particular vendor provided an invoice for payment? (total and incremental)
	Invoice amount
Technology Use	Number of automated actions that were undertaken during this process
Process	Markov method
	Scaled string distance
Outcomes	Length of time spent processing this invoice
	Number of people that 'touched' the invoice

Table 6: Variables contextualized for the invoicing business process

Operationalized Variables

The relationship between characteristics of the invoice and characteristics of the sequences was complex. For example, there was always one entry sequence per invoice, but there could be many approval sequences. The variables that related to the invoice, such as amount and length of time thus could be linked to sequences that also occurred on other invoices. There were many invoices generating the same sequences, and also some sequences appeared on multiple invoices. Table 7 shows how I operationalized the variables, and where each contextualized variable shown in Table 6 fits into the analysis model in Figure 4.

Construct	Variable	Variable Name
Input	Invoice Amount—the log of the invoice amount (per invoice)	LogAvgAmt
	Vendor Count—how many times in total did the vendor have invoices in the entire data set (per invoice)	TotalVendorCount
	Vendor Experience—how many times a particular vendor had invoices prior to the current invoice (per invoice)	VendorExperince
Automation	Number of actions undertaken by the system divided by the total number of actions within a sequence (per sequence)	AutoPCT
Outcome	Length of time between invoice scanning and the completion of the workflow (per invoice)	SC_to_AA

Table 7: Variables operationalized and linked to the analysis model

Invoices sometimes had several amounts that were entered on the system, apparently to update the record as more information was obtained. I used the average of these on each invoice to compute the variable "Invoice Amount". This variable showed a huge range (20 Norwegian kroner (NOK) to 1.5 million NOK), and was skewed left, so the $\log_{10}$ was taken to obtain a more normal distribution (LogAvgAmount).

I used two methods to explore the concept of vendor experience. First, I examined the total number of times that particular vendor appears in all of the data I have (1/1/2002 through 5/31/2006) for each invoice (TotalVendorCount). I also calculated the number of times a particular vendor had appeared on a given invoice before the current invoice (VendorExperience). This second method increases the count by 1 every time an invoice from that vendor occurs. The first invoice from a vendor would have an experience level of 1, the second 2 and so on. These are two similar yet different ways to measure the amount of experience the organization had processing invoices with a particular vendor, and they do correlate highly ($R^2 = .987$, p=0).

Automation was measured at the sequence level, as it was a property of the sequence rather than the invoice. I calculated this variable by dividing the number of automated actions by the total number of actions that comprised the sequence (AutoPCT). This was a relative measure, and thus the percentage of automation can be compared across sequences of different lengths.

Since every invoice in the sample was paid, there was no variance to explain by using this as a measure of outcome. Also, there is an issue here, because 'good' performance of the invoicing process would only be known later, if a correct invoice was not paid or a fraudulent invoice is paid. I chose to use the length of time it took for the organization to approve the invoice as a proxy for how much organizational effort was expended. The actual time used was the number of days it took from when the invoice was scanned to when it was marked 'all approved' or 'fully posted' as noted on Figure 5.

	N	Min	Max	Mean	Std. Deviation
AvgAmount	1990	20.66	1529886	20073.1	74779.704
LogAvgAmount	1990	1.31	6.18	3.6143	.73301
TotalVendorCount	1991	1	2392	516.57	658.595
VendorExperience	2000	0	2117	416.56	552.989
SC_to_AA	1867	0	233	6.17	8.234

Table 8: Descriptive information for invoice variables

Table 8 details some descriptive information for variables that relate to the invoices, while Table 9 shows the information that is based on sequence such as the automation percentage. There were 2000 invoices in the sample that generated 2000 entry sequences and 2852 approval sequences. The majority of the missing N comes from the outcome variable for the length of time for processing. In many of these cases the final date for all-approved was not present, yet the invoice was marked paid. I am assuming there was a problem converting or extracting the dates during the workflow mining phase of the project that lead to this issue.

	N	Min	Max	Mean	Std. Deviation
Entry	2000	0	1.00	.4524	.23241
Approval	2852	0	0.78	.1493	.14698

Table 9: Descriptive information automation percentage

Markov Transition Matrix Method

My first analysis views each sequence as a series of transitions between symbols, based on knowledge of previous symbols in the sequence. In this case, the symbols represent actions that take place within the sequence. I include an overview here but full details and results can be found in chapter 5.

Gottman and Roy (1990) describe a technique using Markov transition matrices to test the effects of contextual variables on the structure of sequences. First, the appropriate model order will be determined by examining the inclusion of additional actions in the sequence, using a nested calculation for entropy. A moving window of varying sizes is passed through the sequence to calculate the amount of information that is gained from knowledge of the differently sized subsequences. A likelihood ratio chi-square test is then performed comparing the hypothesis that the order is r against the null hypothesis that the order is r-1 (Anderson & Goodman, 1957; Gottman & Roy, 1990, p. 62). This tells informs us about the number of subsequences appropriate for further analysis.

Once the order of the Markov process has been determined, the stability of the conditional probabilities between sets of events is measured over time. An omnibus test for stationarity is performed by splitting the sample in half, and a likelihood ratio chi-square test is performed to test for equality of mean probabilities of transition for each sample. Other tests can be performed by splitting the sample into relevant time periods, chosen arbitrarily or as suggested by theory. Trend or cyclic effects can also be tested in this way.

After an examination of temporal stability of the transition probabilities, I investigate the impact of context variables. This is done by splitting the sequences into sets based on these context variables, and testing for homogeneity on the resulting matrices. I split the sample into four groups of similar sizes, stratified by each variable, and test the four groups against each other for homogeneity. The results are detailed in chapter 5.

String Distance Approach

The second analysis I perform uses relative distances between the sequences to obtain a set of metric locations for each sequence in a scaled space. I detail the procedure and the results of the analysis in chapter 6, but I provide a brief overview here.

First, I obtain a string distance (see Appendix 1) for every sequence as it compares to every other sequence in the sample. I chose this distance measure because it has been widely utilized in a variety of fields, and has been used by researchers in the information technology (Sabherwal & Robey, 1993) and organizational routines literatures (Pentland, 2003a, 2003b). String distance also known as Levenshtein distance, has been utilized in sociology and other fields (Abbott, 1990b, 1995; Abbott & Hrycak, 1990; Abbott & Tsay, 2000; Dijkstra, 2001; Dijkstra & Taris, 1995; Sankoff & Kruskal, 1983; van Driel & Oosterveld, 2001) to good result.

The resulting matrix lists each sequence on the column and row headings and the distance between each sequence in the cells. These can be summed across, and a measure of distance can be calculated for each sequence that indicates how different a given sequence is from all of the others (Abbott & Hrycak, 1990; Sabherwal & Robey, 1993). On the other hand, the entire matrix of relational distances can be used as the input to a multidimensional scaling algorithm. Given the ordinal nature of the data, non-metric scaling is most appropriate. This technique extracts the underlying metric relational structure between the sequences within the matrix, based on the structure of their ordinal relationships (Kruskal & Wish, 1978). The extracted dimensions are then examined and explored using multiple correlation and visualization techniques for interpretation. For details and results, see chapter 6.

Chapter 5—Markov Analysis and Findings

Introduction

This chapter describes the first analysis I used to evaluate the relationships between inputs, the process, and outcomes, as well as the impact of technology. I find support for the majority of the model, concluding that vendor experience and invoice amount drive heterogeneity in the process. The entry and approval phases have different relationships with outcomes. Entry processes are homogenous with respect to this variable, but the approval-phase sequences were different when stratified along elapsed time. For both phases, automation emerged as a strong driver of heterogeneity.

After performing the workflow mining discussed in the previous chapter, I convert the sequences into matrices representing the counts of transitions between temporally adjacent actions. An example of this is presented in **Error! Reference source not found.**, Table 32. These matrices are then analyzed by a class of discrete statistics based on the expected and observed probabilities of these transitions, mathematically similar to the analysis of contingency tables (Anderson, 1957; Chatfield 1973, Bishop, Feinberg, and Holland, 1975; Gottman and Roy, 1990).

Determining the Order of the Processes

Following Gottman and Roy (1990), the order of the Markov sequence must be determined at the outset of the analysis. This helps the researcher understand the temporal structure that is present in the set of sequences based on the conditional probabilities of prior steps. According to the likelihood ratio tests suggested by Chatfield (1973), both phases have a digram and trigram structure (**Error! Reference source not found.**, Table 33 and Table 34). This means that the information given by the previous symbol helps predict the current symbol for a digram structure, and the process is of first order. A trigram structure includes information given by the previous two symbols to predict the current symbol and that the process would also be of second order. This does not mean that two or three symbols explain the actual temporal structure, but given the data, those structures best predict the transitions between actions.

Higher-order (second order, third order, etc) Markov transition matrices are often 'sparse', meaning that much of the matrix is 0 or of low observation. This can cause problems for testing the stationarity of the sequence (Gottman and Roy, 1990; Capella 1980). I also examine the scree plots of $\hat{H}_i$ (**Error! Reference source not found.**, Figure 15 and Figure 16), and determine that the additional information from the trigram structure may not be as great as indicated by the likelihood ratio tests (Chatfield 1973, p. 16-17). $\hat{H}_i$ is the amount of information given by the sequence, based on a moving window of size *i*. The scree plot shows that there is not much benefit from moving beyond a trigram structure, but the digram structure still shows an improvement over no temporal information. For mathematic efficiency, I chose to view the entry and approval phase sequences as having digram structure and have first-order Markov properties for the remainder of the analysis.

Examining Stationarity of the Process

I test whether the processes are stable over time, to determine if the transition probabilities change between time periods. The statistical test for this gives a binary result: either the processes are similar (H0) or they are significantly different (Ha). Gottman and Roy (1990) caution readers not to view this omnibus test as an evaluation of the validity of future tests.

I conclude that each set of sequences is not stable over time, according to the omnibus test of stationarity, (Entry Phase p(LR = 2664.487, df =255) = 0, Approval Phase p(LR = 5726.618, df =399) = 0; see **Error! Reference source not found.**, Table 36) Given that the research site is a construction company, there may be a seasonal effect on the processes. It also may be that some of the variables of interest in this study are also moving with time (such as vendor experience), increasing the heterogeneity of the processes along this dimension.

This represents an interesting implication for organizational theory. Organizational routines are widely believed to be rigid and unchanging (M. D. Cohen, 2007), and singular in response to stimuli once search has been eliminated from the process (March & Simon, 1958). A competing perspective highlights organizational routines as a generative structure (Howard-Grenville, 2005; Pentland & Rueter, 1994) - one that generates performances that vary in response to learning (March, 1991; Nelson & Winter, 1982), agency (Feldman, 2000), and is consistent with continuous organizational change (Sorenson, 2003). In a related working paper, Pentland et al (2009b) found this heterogeneity over time in three of four organizations from similar data sets.

I also perform a follow-up test and split the sequences into 4 equal groups to see if I could find some homogeneity over time at a smaller interval than 6 months, and to explore the source of overall difference (**Error! Reference source not found.**, Table 37). For the entry phase, periods 1 and 2 were similar and 3 and 4 were similar, consistent with the omnibus test above, and indicating that there was a natural split at the half-way point of the sample, representing year end. The approval phase showed similarity between the first two periods, and these were distinct from the remaining periods. Periods 3 and 4 were not alike, and distinctly different from 2 and 3. This supports seasonality as a possible explanation for heterogeneity over time in the approval phase.

This finding also connects to the debate between the stable/changing perspectives of organizational routines. As I consider further the groups of process executions (the performative aspects of the organizational routines), I find that there is a natural variation that is present. Also, one would expect that these variations might 'average' themselves out over time, but this is evidence that organizational routines may change over time in a drifting, endogenously changing fashion, rather than an externally driven selection and retention manner. These ideas are developed further in a longitudinal analysis of the performances at four sites in a working paper currently in development (Pentland et al., 2009b).

Group Comparisons

Viewing the processes as sequence with Markov properties allows the comparison of subgroups within the total set of sequences. In addition to time (stationarity), I also evaluate the overall model of input-process-output. Gottman and Roy (1990) suggest segmenting the sequences by variables of interest and deriving the Markov transition matrices of the appropriate order for each segment. These matrices can then be tested with the usual chi-square or likelihood ratio statistics used for contingency tables.

The group comparisons performed here are ones of similarity or differentiation: are groups of processes stratified by a given variable alike or different? Table 10 shows the specific hypotheses that are tested to evaluate the research model, and describes the connections between constructs and variables. All the variables of interest were binned or stratified, with a fixed percentage of cases in each group (25% for four groups). For example, the log of the invoice amounts was calculated for each invoice, and they were ordered smallest to largest. The bottom 25% of the invoices were selected and assigned to group 1. For the entry phase, this means approximately 500 sequences are in each group. The approval phase has more than 500 because there can be many approvals per invoice. The approval phase also did not have exactly the same number of sequences in each group because every invoice did not have the same number of approvals.

The variables measuring input (invoice amount, vendor) and outcome (number of days for the process) are linked to the invoice, while the measure for automation is linked specifically to the sequence. In all cases, the entry and approval sequences are tested separately, but the bins or strata for the variables derived from the invoice are linked to all of the sequences for which that invoice generates. Table 10 on the next page describes these variables, noting which research question and specific hypothesis will be tested in this chapter.

Construct	Variable	Specific Hypotheses
Input	Invoice Amount—the log of the invoice amount (per invoice)	H0: Invoices for large amounts will generate *similar* sequences as those for small amounts Ha: Invoices for large amounts will generate *different* sequences as those for small amounts
	Vendor Count—how many times in total did the vendor have invoices in the entire data set (per invoice)	H0: Invoices from common vendors will generate *similar* sequences as those from uncommon vendors Ha: Invoices from common vendors will generate *different* sequences from uncommon vendors
	Vendor Experience—how many times a particular vendor had invoices prior to the current invoice (per invoice)	H0: Invoices from common vendors will generate *similar* sequences as those from uncommon vendors Ha: Invoices from common vendors will generate *different* sequences from uncommon vendors
Automation	Number of actions undertaken by the system divided by the total number of actions within a sequence (per sequence)	H0: Highly automated sequences will be *similar* to those that are less automated Ha: Highly automated sequences will be *different* from those that are less automated
Outcome	Length of time between invoice scanning and the completion of the workflow (per invoice)	H0: Invoices that take longer to complete will generate *similar* sequences to those that take less time Ha: Invoices that take longer to complete will generate *different* sequences to those that take less time

Table 10: Constructs and variables used to segment sequences

Does the Process Vary with Differential Inputs?

Some organizations have specialized plans in place for different vendors, while others may set up contingent plans to respond to the levels of currency amounts for a particular invoice. In addition, the organization may have routinized responses to vendors it deals most with, so the amount of experience with a particular vendor may affect the process as well.

Phase	Group	N	Avg Log Amt	Min Log Amt	Max Log Amt
Entry	1	498	2.704	1.315	3.080
	2	497	3.364	3.083	3.619
	3	498	3.852	3.621	4.083
	4	497	4.558	4.084	6.185
Approval	1	678	2.717	1.315	3.080
	2	648	3.372	3.083	3.619
	3	657	3.857	3.621	4.083
	4	869	4.591	4.084	6.185

Table 11: Descriptive information for groups of invoice amount, stratified on the log of the amount

First, I examine the impact of invoice total amount on the process (Table 12). Given the large range of currency amounts (20NOK to 1,529,886NOK) and the skewness of its distribution, the $\log_{10}$ was taken of each amount, and this was used to stratify the invoice (Table 11). Table 12 also shows the actual amounts that correspond to the different groups. The likelihood ratio tests indicated that there was heterogeneity in the four groups for both the entry phase and approval phases (Table 13).

Phase	Group	Avg Amt	Min Amt	Max Amt
Entry	1	605.95	20.67	1,201.00
	2	2,460.49	1,211.07	4,162.67
	3	7,451.68	4,176.00	12,095.47
	4	69,838.89	12,128.75	1,529,886.00
Approval	1	613.97	20.67	1,201.00
	2	2,508.92	1,211.07	4,162.67
	3	7,498.74	4,176.00	12,095.47
	4	73,743.40	12,128.75	1,529,886.00

Table 12: Descriptive information for groups of invoice amount, stratified on the log of the amount

For the entry phase, group 1's (20NOK through 1200NOK) processes were different from the remainder. Also, there were no discernable differences in the entry processes that handled the larger invoice amounts (1,200kr thru 1,500,000kr). The approval phase had a similar pattern, but a more complex result. The group with the smallest invoices was different from the rest, but for groups 2, 3 and 4, there were similarities but no transitivity. The conclusion is that the smallest invoices had patterns of action distinct from the remaining groups, and there are some similarities among the groups of sequences of the higher invoice amounts for both the entry and approval phases.

Phase	Test	LR	Df	p-value
Entry	Overall	1323.34	765	0.
	1 v 2	668.39	255	0.
	1 v 3	698.25	255	0.
	1 v 4	859.75	255	0.
	2 v 3	95.47	255	≈1.
	2 v 4	112.69	255	≈1.
	3 v 4	125.16	255	≈1.
Approval	Overall	2015.69	1197	0.
	1 v 2	543.05	399	0.
	1 v 3	844.06	399	0.
	1 v 4	1196.79	399	0.
	2 v 3	233.26	399	≈1.
	2 v 4	468.96	399	0.009
	3 v 4	395.2	399	0.544

Table 13: Group comparisons for invoice amount

I perform several procedures calculating and stratifying the sequences to explore the concept of vendor experience. First, I examine the total number of times that particular vendor appears in all of the data (1/1/2002 through 5/31/2006) for each invoice (Table 14). I also calculate the number of times a particular vendor had appeared on a given invoice before the current invoice (Table 15). This second method increases the count by 1 every time an invoice from that vendor occurs. The first invoice from a vendor would have an experience level of 1, the second 2 and so on. These are two similar ways to measure the amount of experience the organization had processing invoices with a particular vendor, and they correlate highly (R^2 =.987).

Phase	Group	N	Avg of Total Vendor Count	Min	Max
Entry	1	515	26.074	1	66
	2	485	142.054	68	247
	3	533	468.407	250	763
	4	458	1,520.766	772	2392
Approval	1	656	27.407	1	66
	2	788	141.201	68	247
	3	667	456.772	250	763
	4	741	1,682.767	772	2392

Table 14: Groups and descriptives for total vendor count

Phase	Group	N	Avg of Vendor Experience	Min	Max
Entry	1	502	15.787	0	42
	2	498	100.325	43	187
	3	502	330.711	188	546
	4	498	1,223.329	547	2117
Approval	1	612	16.333	0	42
	2	780	97.759	43	187
	3	683	323.316	188	546
	4	777	1,381.834	547	2117

Table 15: Groups and descriptives for vendor experience

Taken together, these two related measures paint a similar overall picture that is different in only a few details (Table 16). The overall tests show that the four groups are heterogeneous in their processes, and for the most part are consistent across subgroup tests. There is some indication that the membership of groups 2 and 3 were different between the two variables for the approval phase. Together, these results suggest that the experience an organization has with a vendor does impact the process for those invoices. The organization has similar processes for invoices from uncommon vendors, and different processes for common vendors, a finding that is consistent with the law of requisite variety.

		Total # of Invoices			Experience with Vendor		
Phase	Test	LR	df	p-value	LR	df	p-value
Entry	Overall	2215.45	765	0.	1878.10	765	0.
	1 v 2	212.21	255	0.976	259.65	255	0.407
	1 v 3	1083.81	255	0.	832.16	255	0.
	1 v 4	1119.26	255	0.	1095.83	255	0.
	2 v 3	746.51	255	0.	426.17	255	0.
	2 v 4	795.5	255	0.	682.57	255	0.
	3 v 4	378.92	255	0.	361.03	255	0.
Approval	Overall	3313.87	1197	0.	3182.70	1197	0.
	1 v 2	383.96	399	0.697	437.78	399	0.088
	1 v 3	486.38	399	0.002	504.65	399	0.
	1 v 4	1481.98	399	0.	1476.87	399	0.
	2 v 3	485.69	399	0.002	387.19	399	0.655
	2 v 4	1642.83	399	0.	1517.57	399	0.
	3 v 4	1382.16	399	0.	1266.63	399	0.

Table 16: Group comparisons for total number of invoices for a vendor and the vendor experience, for each invoice.

This analysis does not allow one to discern what exactly the differences between sets of process executions are. The use of order statistics, introduced in the discussion section, would allow the extraction of a 'primal' pattern of each set of invoices, and these differences between common vendor patterns and singular vendor patterns could be enumerated. This would be an interesting extension of this dissertation.

Are Processes that Have Different Outcomes Similar?

The next step was to determine whether processes that had similar outcomes were generated by similar processes. I calculated the time that a particular invoice spent in process within the workflow—from the time it was scanned to when it was flagged 'all-approved'. Some of the invoices had missing data for either the start or end date, and the length of time for the process was incalculable. Table 17 shows these and the characteristics of the groups stratified by process time.

Phase	Group	N	Avg of Days of Process	Min	Max
Entry	Missing	133			
	1	576	1.286	0	2
	2	487	4.061	3	5
	3	419	6.899	6	8
	4	385	15.366	9	233
Approval	Missing	326			
	1	779	1.287	0	2
	2	654	4.013	3	5
	3	591	6.90	6	8
	4	502	15.09	9	233

Table 17: Descriptives for the length (in days) of the process

There was a marked difference in how the processes related to outcome between the entry and approval phases (Table 18). When the entry phase processes were stratified according to the length of time the invoice spent in the workflow, they were homogeneous overall. The approval phase processes were different from each other, even with the same categorization scheme as the entry phase processes. This may be because much of the work that takes place during the time elapsed between scanning and approval takes place during the approval phase. The entry phase would almost never take more than one day on its own.

Phase	Test	LR	df	p-value
Entry	Overall	686.71	765	0.98
	1 v 2	168.19	255	1.
	1 v 3	188.66	255	0.999
	1 v 4	371.72	255	0.
	2 v 3	134.92	255	1.
	2 v 4	290.21	255	0.064
	3 v 4	181.91	255	1.
Approval	Overall	2149.54	1197	0.
	1 v 2	653.23	399	0.
	1 v 3	1100.11	399	0.
	1 v 4	1386.22	399	0.
	2 v 3	264.61	399	1.
	2 v 4	476.79	399	0.004
	3 v 4	320.54	399	0.999

Table 18: Comparing processes with similar outcomes

The subgroup tests suggest a much more complicated story. Even though the sequences were similar overall, groups 1 and 4 were different enough in the entry phase to become significant. The approval phase subtests shows a very interesting result, where periods 2 and 3 were similar along with 3 and 4, yet period 2 was different from period 4. I believe that transitivity may not apply to the tests of homogeneity for these processes, or that 2 is 'enough' like 3 and 3 is 'enough' like 4 to pass the test, yet 2 and 4 are different 'enough' in transition probabilities to be significantly different when stratified by the elapsed time of the process.

This brings to mind the paradox of the 'ship of Theseus', and is related to similar discussions in philosophy of change and identity. In the legend, as pieces of the ship wore out, they were replaced with new timber and planks to the point where none of the original pieces were present in the ship. The question arises, is this still the ship of Theseus? The transitivity between processes stratified by their outcome brings this to the fore, even though time is not the dimension that change is measured against. At one

level, it is all invoicing, and almost all the invoices are processed within 15 days of scanning them. The Markov method allows me to discern when processes that I 'identify' as the same exhibit differences in their sequence and choice of actions, and conclude that there are differences at a lower level of abstraction and higher level of detail. Since the processing time of invoices may be variable of interest for managers seeking to control costs, the issue of transitivity and the question of 'same or different' represent another significant contribution of this work.

Does the Process Vary with the Amount of Automation Present?

The amount of automation present in a given sequence is measured as the number of actions undertaken by the workflow system in a sequence divided by the length of that sequence. This measure is related to the sequence, rather than the invoice. Thus, a given invoice may have an automation score for the entry phase, and several different scores for each of the approval sequences that are present. Table 19 and Table 20 show these groups and their characteristics, including the number of automated actions, the length of sequences, and finally the ratio of automated actions to sequence length.

Phase	Group	N	Avg Auto Actions	Min	Max	Avg Length	Min	Max
Entry	1	551	1.89	0	2	10.26	9	14
	2	469	3.64	3	5	10.15	9	20
	3	596	5.52	5	6	10.14	10	14
	4	384	8.20	6	10	10.12	9	13
Approval	1	620	0.29	0	3	11.22	2	45
	2	924	1.53	1	4	13.63	7	31
	3	607	2.17	1	4	13.44	6	27
	4	701	4.74	1	13	15.21	4	37

Table 19: Groups and descriptives for automation (automated actions, sequence length).

Phase	Group	N	Avg Auto %	Min	Max
Entry	1	551	18.43%	0.%	20.00%
	2	469	35.92%	25.00%	40.00%
	3	596	54.48%	41.67%	60.00%
	4	384	81.06%	63.64%	100.00%
Approval	1	620	1.96%	0.%	7.41%
	2	924	10.97%	7.69%	14.29%
	3	607	16.14%	14.81%	18.18%
	4	701	31.95%	18.75%	78.57%

Table 20: Groups and descriptives for automation (automation percent).

Phase	Test	LR	df	p-value
Entry	Overall	11625.23	765	0.
	1 v 2	3557.55	255	0.
	1 v 3	5901.05	255	0.
	1 v 4	3797.48	255	0.
	2 v 3	1845.35	255	0.
	2 v 4	2806.09	255	0.
	3 v 4	2975.18	255	0.
Approval	Overall	5532.13	1197	0.
	1 v 2	1892.11	399	0.
	1 v 3	2660.79	399	0.
	1 v 4	3189.06	399	0.
	2 v 3	770.61	399	0.
	2 v 4	1452.01	399	0.
	3 v 4	998.02	399	0.

Table 21: Group comparisons for the amount of automation present within a process

The results of the likelihood ratio tests indicate that for each phase, each of the four stratified groups is different in their processes (Table 21). I explore this further by examining histograms, and then use an alternate stratification process. Figures 6 and 7 show the average amount and distribution of automation percentage is very different between phases.

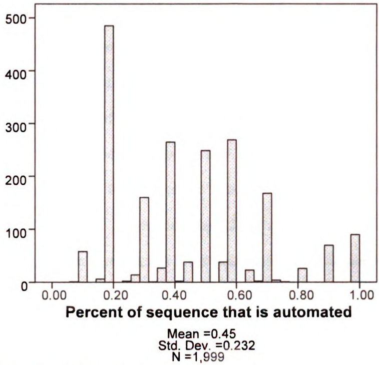


Figure 6: Distribution of automation for entry phase.

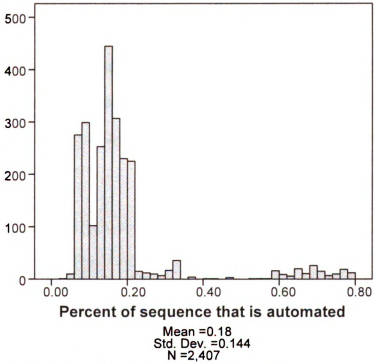


Figure 7: Distribution of automation for approval phase

The approval phase is marked by a smaller overall amount of automated actions based on mode and mean. I also bin the approval phase into chunks that would separate out the modal set from the grouping of processes at .55 automation and higher obtaining similar results. Instead of stratifying based on quartiles, I perform a visual stratification technique based on their distribution (figures 6 and 7). There were still four groups (**Error! Reference source not found.**, Table 35), but I capture modal areas together instead of groups of equal size. As seen in Table 22, the results are qualitatively identical to those in Table 21. Taken together, this gives strong evidence that processes that are more highly automated have patterns of action that are distinct from those with less automation.

The implication is that automation affects the temporal and task structure that digitally-enabled organizational routines exhibit. This is consistent with the literature on technology impact, but the analysis performed here represents a new way of looking at technology and its effect on business processes. This allows a much finer grained measure of technology use, in the percent of each execution that is automated. The use of order statistics as an adjunct to this analysis could allow the extension of this and allow further discovery.

Phase	Test	LR	df	p-value
Entry	Overall	9150.95	765	0.
	1 v 2	935.93	255	0.
	1 v 3	2417.66	255	0.
	1 v 4	2705.73	255	0.
	2 v 3	3493.46	255	0.
	2 v 4	4508.71	255	0.
	3 v 4	2380.91	255	0.
Approval	Overall	7064.57	1197	0.
	1 v 2	2354.58	399	0.
	1 v 3	2863.56	399	0.
	1 v 4	4276.91	399	0.
	2 v 3	655.23	399	0.
	2 v 4	1324.42	399	0.
	3 v 4	1242.19	399	0.

Table 22: Group comparisons for the amount of automation present within a process, visual stratification process

Discussion

When a routine is viewed as a Markov process, there are several connections to how organizational routines are theorized. The probabilities of transition between actions within a sequence can be thought of as representing the dispositions or habitual nature of routines (Pentland et al., 2009d). This makes the Markov approach especially useful to study what the performances are, how many different types there are, and how changes unfold. This perspective also fits with the theoretical concepts of patterns and essential variety that are found in the performance of organizational routines.

Tests of homogeneity allow the researcher to discover how alike or different two sets of performances are from each other, and may lead to changes in how we identify routines as the same or different, or know if we have one routine or many. The discovery of similarities and differences between performance sets that were not transitive was unexpected. This may challenge how we think of the concept of routine identity.

The concept of stationarity is similar to our understanding of how changes in organizational routines can be seen in situ. Routines that exhibit change, even incremental change, can be seen as alterations in the choice and temporal structure of actions. The order of the process can be seen as the amount of temporal interdependence between actions, but limitations from the sparseness of the transition matrix make this difficult mathematically to apply to the data collected for this dissertation.

While the analysis employed here only allows the investigation into similarity or difference as a binary decision, the overall research model was well supported (Table 23). There are similarities between processes with similar inputs (no group 1 was statistically similar to any group 4). For the approval phase, there are differences between processes with differential outcomes. This makes sense, since the majority of the 'work' of approving an invoice does not take place during the entry phase.

Automation seems to drive heterogeneity in the process, in that when the processes were stratified by the percent of actions that were automated, no group was similar in transitions to any other. On the other hand, this may be an indication of the endogeneity of technology within the process—actions are automated with different frequency because those tasks are easier or more difficult for the system to do them without human interaction. In the next chapter, I complement this analysis with the results from the string matching analysis.

<table>
<tr><th>Research Question</th><th>Variable</th><th>Results</th></tr>
<tr><td rowspan="6">Does the process vary with differential inputs?</td><td rowspan="2">Invoice Amount—the log of the invoice amount (per invoice)</td><td>Entry—Group 1 is different from groups 2, 3, 4
<table>
<tr><td></td><td>1</td><td>2</td><td>3</td><td>4</td></tr>
<tr><td>1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>2</td><td></td><td></td><td>●</td><td>●</td></tr>
<tr><td>3</td><td></td><td>●</td><td></td><td>●</td></tr>
<tr><td>4</td><td></td><td>●</td><td>●</td><td></td></tr>
</table></td></tr>
<tr><td>Approval—
<table>
<tr><td></td><td>1</td><td>2</td><td>3</td><td>4</td></tr>
<tr><td>1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>2</td><td></td><td></td><td>●</td><td></td></tr>
<tr><td>3</td><td></td><td>●</td><td></td><td>●</td></tr>
<tr><td>4</td><td></td><td></td><td>●</td><td></td></tr>
</table></td></tr>
<tr><td rowspan="2">Vendor Count—how many times in total did the vendor have invoices in the entire data set (per invoice)</td><td>Entry— Groups 1 and 2 are different from 3 and 4
<table>
<tr><td></td><td>1</td><td>2</td><td>3</td><td>4</td></tr>
<tr><td>1</td><td></td><td>●</td><td></td><td></td></tr>
<tr><td>2</td><td>●</td><td></td><td></td><td></td></tr>
<tr><td>3</td><td></td><td></td><td></td><td></td></tr>
<tr><td>4</td><td></td><td></td><td></td><td></td></tr>
</table></td></tr>
<tr><td>Approval—
<table>
<tr><td></td><td>1</td><td>2</td><td>3</td><td>4</td></tr>
<tr><td>1</td><td></td><td>●</td><td></td><td></td></tr>
<tr><td>2</td><td>●</td><td></td><td>●</td><td></td></tr>
<tr><td>3</td><td></td><td>●</td><td></td><td></td></tr>
<tr><td>4</td><td></td><td></td><td></td><td></td></tr>
</table></td></tr>
<tr><td rowspan="2">Vendor Experience—how many times a particular vendor had invoices prior to the current invoice (per invoice)</td><td>Entry—
<table>
<tr><td></td><td>1</td><td>2</td><td>3</td><td>4</td></tr>
<tr><td>1</td><td></td><td>●</td><td></td><td></td></tr>
<tr><td>2</td><td>●</td><td></td><td></td><td></td></tr>
<tr><td>3</td><td></td><td></td><td></td><td></td></tr>
<tr><td>4</td><td></td><td></td><td></td><td></td></tr>
</table></td></tr>
<tr><td>Approval—
<table>
<tr><td></td><td>1</td><td>2</td><td>3</td><td>4</td></tr>
<tr><td>1</td><td></td><td>●</td><td></td><td></td></tr>
<tr><td>2</td><td>●</td><td></td><td>●</td><td></td></tr>
<tr><td>3</td><td></td><td>●</td><td></td><td></td></tr>
<tr><td>4</td><td></td><td></td><td></td><td></td></tr>
</table></td></tr>
<tr><td rowspan="2">Does the process vary with the amount of automation present?</td><td rowspan="2">Number of actions undertaken by the system divided by the total number of actions within a sequence (per sequence)</td><td>Entry—All strata are different</td></tr>
<tr><td>Approval— All strata are different</td></tr>
</table>

Table 23: Summary of results—bullets indicate similarity, or an insignificant statistical test

<table>
<tr><td rowspan="12">Are processes that have different outcomes similar?</td><td rowspan="12">Length of time between invoice scanning and the completion of the workflow (per invoice)</td><td colspan="5">Entry—</td></tr>
<tr><td></td><td>1</td><td>2</td><td>3</td><td>4</td></tr>
<tr><td>1</td><td></td><td>●</td><td>●</td><td></td></tr>
<tr><td>2</td><td>●</td><td></td><td>●</td><td>●</td></tr>
<tr><td>3</td><td>●</td><td>●</td><td></td><td>●</td></tr>
<tr><td>4</td><td></td><td>●</td><td>●</td><td></td></tr>
<tr><td colspan="5">Approval—</td></tr>
<tr><td></td><td>1</td><td>2</td><td>3</td><td>4</td></tr>
<tr><td>1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>2</td><td></td><td></td><td>●</td><td></td></tr>
<tr><td>3</td><td></td><td>●</td><td></td><td>●</td></tr>
<tr><td>4</td><td></td><td></td><td>●</td><td></td></tr>
</table>

Table 23 Continued

Chapter 6—String Distance Analysis and Findings

Introduction

In this chapter, I describe the analysis utilizing string distance and multidimensional scaling techniques to explore how inputs drive sequential variety and its impact on outcomes. I also examine the impact of automation on the process. This second method views each process execution as an ordered set of symbols, and calculates distances between them based on the number of insertions, deletions and substitutions of symbols that it takes to convert one sequence into another. I begin this chapter with a description of how I prepared the data for analysis. Then, I discuss multidimensional scaling as a way to extract the relationships between sequences, and then analyze these relationships by using a partial least squares technique.

Overall, I find support for the research model, but there are also differences between the entry and approval phases of the invoicing business process in the nature and form of both buffering and automation. A key finding is the differing role of automation between the entry and approval phases in that automation serves as a substitute for action-based buffering in the entry phase, and a complement in the approval phase. Using scaled string distance holds promise for empirically studying organizational routines. With this data, the method correctly discriminates between types of sequences from entry and approval, and extracts variables that describe facets of the process.

Despite the confirmatory nature of the research design, the application of MDS to string distance in this dissertation represents an exploration. Given idiosyncrasies of the data and the structure of the model, PLS was not the best choice for analysis. This data will be reanalyzed using GLM, logit or probit to deal with problems of non-normality, identification and endogeneity. Initial work with these methods shows promise and may lead to more robust findings than those presented here.

Model Variables

The invoice parameters are inputs to the invoice handling process, and the process generates outputs and outcomes. I am using the invoice amount (LogAvgAmt) and the experience that the organization has with the invoice vendor (TotalVendorCount, VendorExperience) as the input variables. Because of the extremely skewed distribution of invoice amount, I use the $\log_{10}$ instead of the raw amount. As with the Markov analyses, automation percent (AutomationPCT) is a ratio of the number of automated actions within a sequence divided by the length of the sequence. Outcomes are defined as the length of time spent processing from data entry to final approval (SC_to_AA). The main difference in this analysis from the Markov analysis is in the way I represent the sequence in the analysis, and the techniques I am using to model my theory.

Preparing the Data

As with the Markov analyses, I process the event log into a set of sequences related to each invoice. This results in a list of entry sequences (one for every invoice) and a list of approval sequences. I then remove the duplicate sequences from the set leaving a list of sequence types for both entry and approval. This is done because the multidimensional scaling algorithms I employ do not support having items with zero distance in the matrix.

Then, I create an N by N matrix, N equaling the number of sequence types in the set. Each cell in the matrix represents the number of insertions, deletions and substitutions that at takes to convert one sequence into the other, or the string distance between each sequence. This measure, known also as a Levenshtein distance (Sankoff & Kruskal, 1983), has been utilized in studies found in various literatures (Abbott, 1983, 1990b, 1995; Sabherwal & Robey, 1993). I chose this distance measure because it has been used in the realm of information systems development (Sabherwal & Robey, 1993), was suggested for use by social science researchers (Abbott, 1983, 1990a, 1995), and specifically has been used to describe the concept of sequential variety in organizational routines (Pentland, 2003a, 2003b). Another commonly used distance measure is cosine angle distance, typically used for interpretation of search queries and AI language processing, but this doesn't have the rich theoretical connections of the Levenshtein distance.

This matrix is considered a dissimilarity matrix, as a zero represents identity, and higher numbers indicate sequences that are more dissimilar. These relationships between sequences are ordinal, rather than metric, suggesting non-metric multidimensional scaling should be utilized. (Kruskal & Wish, 1978).

Initial Exploration

I performed several analyses with various scaling algorithms and techniques, starting with the same initial dissimilarity matrix. I knew from the system and the data that there were two phases of the invoice approval process. I decided to put both sets of sequences into the same matrix and determine where the scaling would locate them. This is shown in Figure 8.

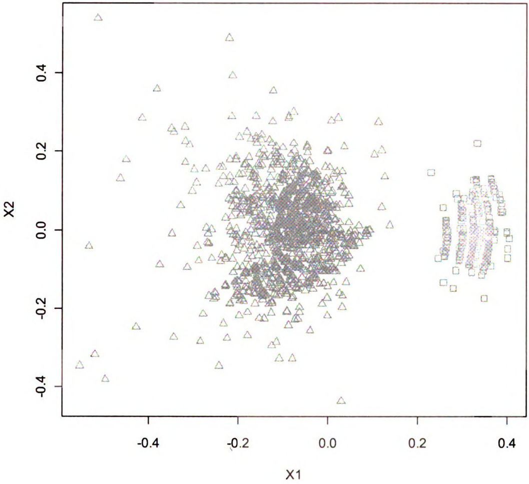


Figure 8: Entry (square) and approval (triangle) scaled in two dimensions.

For this application, two dimensions of scaled string distance clearly group all of the entry sequences (square) together, and approval sequences (triangle) together in another location. Even if it were not known that there were two groups of sequences, this analysis would have indicated this heterogeneity. Different subroutines within an overall routine are discerned, an important demonstration of this technique when applied to performances of organizational routines. Also, note that the entry sequence cluster is much more compact than the approval sequence cluster. This is an indication that the entry sequences are more homogenous than the approval sequences, and is consistent with sequential variety measures for each group (Pentland et al., 2009d).

Extracting Metric Relationships between Sequences

To perform multidimensional scaling, some choices are required by the researcher, based on the purpose of the analysis and idiosyncrasies of the data. First, there are several algorithms available to convert the dissimilarity matrix into a distance matrix. Second, the number of dimensions that the underlying data will be projected onto must be determined. Two criteria are suggested to help the researcher make these decisions, one based on a fit measure, the other on the pragmatic use of the dimensional data.

$$S = \sqrt{\frac{\Sigma_{i \neq j}\left[\theta\left(d_{ij}\right) - \tilde{d}_{ij}\right]^2}{\Sigma_{i \neq j}\tilde{d}_{ij}^{\,2}}}$$

Equation 1: Stress

When the non-metric multidimensional scaling (NMDS) algorithm converges, it gives a measure related to the fit of the solution called 'stress' (equation 1). A researcher should seek to minimize the stress while preserving his or her ability to understand or use the data in analysis. Stress, then is a measure of 'badness of fit'(Kruskal & Carrol, 1969; Kruskal & Wish, 1978). Two sets of visualization plots help make this decision, along with the calculated value of stress. A Sheppard plot is used to graph ordination distances against original dissimilarities, and also gives a goodness of fit measure (Oksanen, 2009b).

Choosing a Distance Function

I decided to examine some other fields and how they utilize multidimensional scaling. Vegetation ecology researchers use non-metric multidimensional scaling techniques to understand the relationships between the ecological content of a given area and variables such as altitude, orientation, rainfall, and soil composition. Researchers in this field suggest that Euclidian and Jaccard distance functions tend to do the best job of extracting meaningful dimensions from data similar to mine (Oksanen, 2009a). The statistical package I used in the R program was developed specifically for vegetation ecologists, but the MDS functions are generic and have been used by researchers in other fields as well (Oksanen, 2009a).

$$d_{jk} = \sqrt{\sum_{i=1}^{N} (x_{ij} - x_{ik})^2}$$

Equation 2: Euclidian Distance Calculation

I evaluate a non-transformational conversion (raw), one based on Euclidian distance (Equation 2) and one based on Jaccard distances (Appendix 3, Equation 5). The raw distance matrix is simply the matrix of string distances between sequences. Figure 9 and Figure 10 show a scree plot of the stress for each of the distance functions, for extracting dimensions 1 through 5. The shape of the curve is diagnostic, and should be monotonically downward sloping. Figure 10 shows that the approval phase raw distance data caused some difficulty for the MDS algorithm when moving from 2 to 3 dimensions. In both the entry and approval phases, the Jaccard and Euclidian distance functions performed better than the raw distance, and there was little difference between the absolute stress levels for these two distance calculations.

Error! Reference source not found. in Appendix 3 show the Sheppard plot for the MDS projections of the Euclidian distance matrix and 1, 2, 3, 4, and 5 dimensions. I also evaluated the Sheppard plots of the Jaccard and raw matrix MDS solutions, but these are not shown. Given the higher stress values for the raw matrix, and the lack of significant difference between Jaccard and Euclidian for the scree plot and Sheppard diagrams, I chose to use the Euclidian distance for the remainder of the analysis.

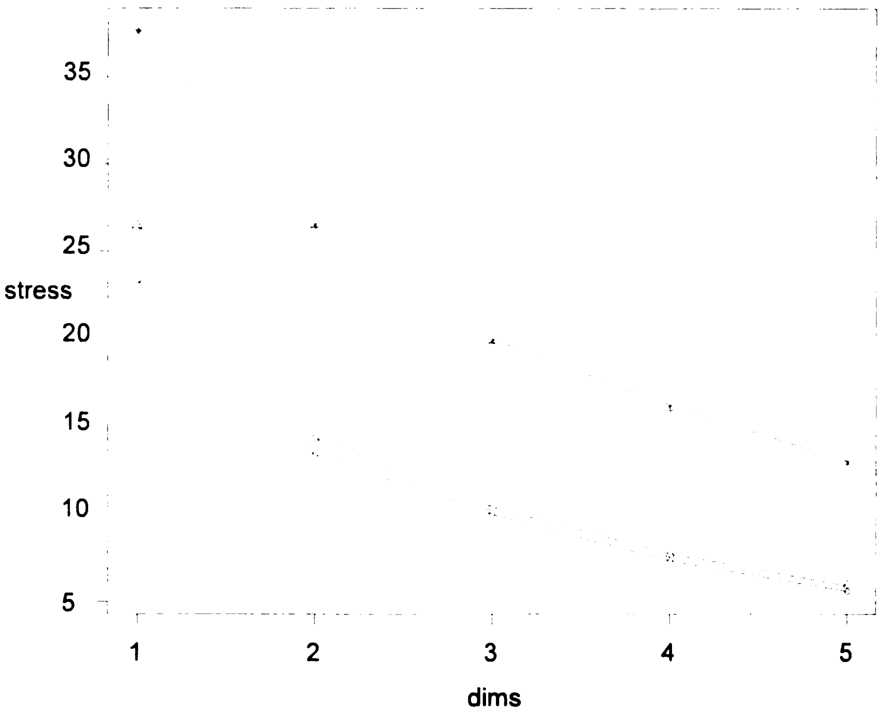


Figure 9: Entry scree plot—Top line is raw, next line down is Jaccard, bottom line is Euclidian. Stress is the y-axis, number of dimensions on the x-axis

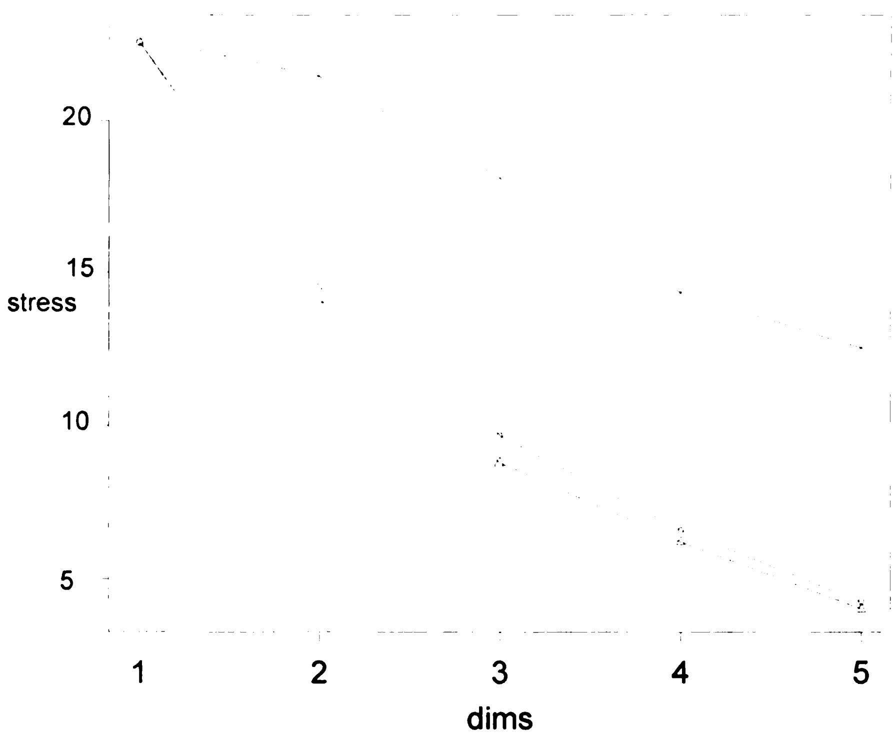


Figure 10: Approval scree plot—Top line is raw, next line down is Euclidian, bottom line is Jaccard. Stress is the y-axis, number of dimensions on the x-axis

Determining the Appropriate Number of Dimensions

Examining a scree plot of the stress over multiple dimensions is also helpful in deciding how many should be utilized. As seen in Figure 9, it appears that a two or three dimension solution would be acceptable for the entry phase and the marginal benefit of moving to four or more dimensions may be low. The approval phase scree plot (Figure 10) indicates that three or four dimensions may be needed to ensure a better fit of the data with the projection.

In many cases, the guide for choosing dimensionality should be the pragmatically driven by the interpretability of the solution (Kruskal & Wish, 1978). This involves examining several projections and visualizing their locations in space, running multiple regressions and evaluating how well the dimensions explain the data. The three-dimensional projections can be viewed two dimensions at a time, or visualized interactively in a 3-d plot. Visualization beyond three dimensions is difficult.

Results

In the next section, I discuss the results of the analyses performed in this chapter. I first qualitatively examine a list of sequences and the dimensions obtained from the MDS algorithm. Then, I use multiple-regression to help interpret these dimensions in relation to the other variables in my model. These regressions also can be used to visualize how the MDS dimensions relate to the input, outcome, and automation variables. Next, I employ these dimensions as formative measures of the process it in a partial-least squares structure using smartPLS. Finally, I discuss the relationships between inputs, process, outcome and automation based on this path analysis.

Interpreting the MDS Sequence Dimensions

To complement the Markov analysis, I attempt to discover which input variables drive differences in the process. Kruskal and Wish (1978) suggest several techniques to help the researcher understand what the variables from a specific MDS projection mean. First, they suggest considering the dimensions and relate them to the original data. In my case, this means qualitatively examining a list of sequences with each sequence's scores on each of the extracted dimensions. A variety of visualizations, including plotting the points representing the sequences and coloring them in relation to other variables can also be helpful. Finally, they suggest multiple-regression of projected MDS dimensions on each variable of interest to evaluate the significance and explained variance of a given solution to evaluate how they relate. Kruskal and Wish (1978) imply candidate covariates should have as high an R^2 as possible and indicate a .01 alpha level for testing regression significance. They give a rule of thumb at .7 for acceptable explained variance (R^2), but they note that this is not always possible in practice.

Qualitative Analysis of MDS Dimensions

I place the sequences in a list with the related dimensions that are extracted for each sequence. I evaluate sequence length, and I discover that there appear to be no relationship for the entry phase, but v2 on the approval phase indicates a noisy relationship with sequence length. I then sort the sequences by each dimension in order to determine if I could discern any additional patterns. This was difficult in some cases, because the underlying relationships may be at an angle to these listed dimensions. There are an infinite number of unique projections that show the same structure from the MDS algorithm, all rotations along some axis from each other. Table 24 indicates some of the regularities that I was able to observe in these sequences, and what those action sequences actually are.

Entry Phase	V1	High values seem to start with 5, 8, 7, 6 more often than lower values. This is a sequence of activities "Enter document type", "Enter invoicno", "Enter invoicedate", "Enter duedate". Lower values did not seem to have a common pattern.
	V2	Higher values seemed to start with 8,7,6,3,4 often with the action sequence: "Enter invoicno.", "Enter invoicedate", "Enter duedate", "Enter amount", "Enter currency". Lower values seemed to start with 5, 4, 20 and 5, 3, 4 often giving the action sequence: "Enter document type", "Enter currency", "Enter vendor account", "Enter value dim. 7".
	V3	I did not identify any regularities.
Approval Phase	V1	Loosely affiliated with sequence length. Idid not identify any differences between high and low values.
	V2	Fairly good relationship with sequence length. Low values (< -35) tended to start with 2, 11, 10, or an action sequence of "Enter account", "Enter Tax-code", "Enter period". Higher values (>-35 and < 40) were more likely to start 10, 4, 23 for an action sequence of "Enter period", "Enter amount", "Enter account".
	V3	I did not identify any regularities here.

Table 24: Qualitative results from examining raw sequences and extracted variables

Implications of Sample Size

In Table 25, I show the number of sequences, the number of sequence types, and the number of invoices that are the basis for the number of sequences. I sampled 2000 invoices, but this leads to 2000 entry and 2853 approval sequences. Of these sequences, there were 206 entry phase types and 929 approval phase types. There were some missing values in the outcome variable (SC_to_AA), so the number of valid N is smaller than this number. Variables related to the invoice such as those related to inputs (TotalVendorCount, VendorExperience, LogAvgAmount) and those related to outcomes (SC_to_AA) are based on the invoice. The MDS variables locating the sequences by their scaled string distance are based on the types of sequences. Automation is measured separately for each sequence, so there are as many different values for this as there are actual valid sequences.

Phase	Invoice Data	Number of Sequences	Sequence Types	Valid N	Automation Data
Entry	2000	2000	206	1869	1869
Approval	2000	2853	929	2528	2528

Table 25: Sample size and basis for samples and data

The total information (variance) that was available in the sample was less than the total amount that was possible (e.g., 206 sequence types in 2000 sampled sequences). Given the lexicon and length of sequences, the reduction in information content is quite large. The total number of possible sequences in this system is truly infinite, because there is no upper limit on the length of the sequence, and the lexicon is editable. When r-square is calculated, it is the ratio of explained variance to the total variance (explained + unexplained). I believe that the unexplained variance is inflated for one main reason: redundancy of sequence types in my sample

I could at most find 2000 different entry sequences and 2853 approval sequences. I actually found 206 different entry and 908 approval sequences. This means that the MDS algorithm only had a small set of different string distances as compared to the variance that could be observed. I believe this smaller set is the actual amount of unexplained variance relating to the location of the sequences in space. I can only explain at most the variance based on 206 sequences in the approval phase, even though my N is based on a sample of 2000. Simply adjusting N to 206 or 929 will not solve the problem, as I really do have samples of 2000 and 2853 for entry and approval respectively. As I note in the discussion of this chapter, there is a calculable correction for this, but this was not performed.

	N	Min	Max	Mean	Std. Deviation
AvgAmount	1990	20.66	1529886	20073.1	74779.704
LogAvgAmount	1990	1.31	6.18	3.6143	.73301
TotalVendorCount	1991	1	2392	516.57	658.595
VendorExperience	2000	0	2117	416.56	552.989
SC_to_AA	1867	0	233	6.17	8.234

Table 26: Descriptive information for invoice variables

	N	Min	Max	Mean	Std. Deviation
Entry	2000	.00	1.00	.4524	.23241
Approval	2852	0	0.78	.1493	.14698

Table 27 Descriptive information for sequence variables

Using Multiple Regression to Understand MDS Dimensions

Tables 26 and 27 above show the descriptive information for the variables relating inputs, outcomes, and automation. I evaluate each of the distance metrics for two and three dimensional projections by performing a regression with the variable of interest as dependent variable, and the MDS dimensions as independent variable. This was done for both the entry phase and the approval phase. The coefficients can then be used to map a line or 'gradient' upon the scatterplot of scaled sequences to show the direction that variable moves within sequence space as a way to visualize these relationships. This is seen in the Appendix 3, figures 18 through 21. Table 28, Table 29 and Table 30 show regressions of each variable of interest upon the two and three-dimensional MDS

Euclidian solution. Each table focuses on a separate set of variables relating to inputs, automation, and outcome.

	Phase	Variable	Z.v1	Z.v2	Z.v3	r-Square	Adj. r-Square
Environment	Entry	Vendor Experience	-0.096*	0.291*		0.089	0.088
			-0.022*	0.299*	-0.265*	0.159	0.157
		Log Amount	0.086*	-0.174*		0.034	0.033
			0.087*	-0.180*	-0.054*	0.038	0.037
	Approval	Vendor Experience	-0.090*	-0.275*		0.090	0.089
			-0.093*	-0.260*	-0.079*	0.096	0.095
		Log Amount	0	0.006		0.071	0.005
			0.002	0.094*	-0.121*	0.019	0.018

Table 28: Two and three dimensional standardized regressions on input variables, entry and approval phase. * indicates significance at the .01 level

For entry phase, the two dimensional sequence space explains roughly 9% of the variance in vendor experience, and this rises to 15.7% when the third extracted dimension is included, as seen in Table 28. Similar results were initially seen for the approval phase, adding a third dimension did not add much explanatory power over the two dimensional solution. The log of the vendor amount was explained weakly by the extracted variables for the entry phase, but the approval phase indicated a poor fit and low explanation.

	Phase	Variable	Z.v1	Z.v2	Z.v3	r-Square	Adj. r-Square
Outcome	Entry	SC_to_AA	-0.265*	0.155*		0.068	0.067
			-0.258*	0.154*	-0.014	0.067	0.066
	Approval	SC_to_AA	-0.223*	0.118*		0.054	0.053
			-0.220*	0.109*	0.054*	0.056	0.055

Table 29: Two and three dimensional standardized regressions on outcome variables, entry and approval phase. * indicates significance at the .01 level

Table 29 interprets how well the extracted dimensions explain the outcome of the process, the amount of time it takes from scanning to approval (SC_to_AA). There is a good pattern of significance overall, and the r-square indicates a weak relationship between the process variables and outcomes. Interestingly, adding a third dimension to the process added little explanatory power as evidenced by no significant increase in adjusted r-square.

	Variable	Phase	Z.v1	Z.v2	Z.v3	r-Square	Adj. r-Square
Automation	Entry	AutoPCT	-0.270*	0.333*		0.169	0.168
			-0.103*	0.365*	-0.645*	0.570	0.569
	Approval	AutoPCT	-0.197*	-0.002		0.039	0.038
			-0.194*	-0.023*	0.109*	0.051	0.050

Table 30: Two and three dimensional standardized regressions on automation variables, entry and approval phase. * indicates significance at the .01 level

Automation seems to be the variable (of those tested) that is best explained by the location of sequences in multidimensional scaled space. The three-dimensional solutions explain more variance in the amount of automation than does the two-dimensional projection for the entry phase (Table 30). While an R^2 of .569 does not meet Kruskal's rule of thumb of .7, this is the best result obtained, suggesting that the three-dimensional projection of automation onto the scaled space best explains the variance in the process. The same result was not seen for the approval phase where the amount of explained variance was low, and did not improve much by including an additional dimension.

I summarize the differences in adjusted R^2 due to the addition of the third extracted dimension in Table 31 for all of the variables evaluated in Table 28, Table 29 and Table 30. Because there was at least some improvement in most cases, I decided to use the three-dimensional scaling solution for visualization and path analysis.

	Variable	Entry	Approval
Environment	Vendor Experience	0.069	0.006
	Log Amount	0.004	0.013
Outcome	SC_to_AA	-0.001	0.002
Technology Use	AutoPCT	0.401	0.012

Table 31: Improvement in explained variance (adjusted r-square) in 3-dimensional solution over the 2-dimensional solution

Multiple Regression and Visualization

I examined visualizations of the projected dimensions, along with the regressions of these dimensions on the other variables of interest. It is difficult to show a vector or surface for how the variables relate to the 'cloud' of points representing the sequences, at least on paper. Sometimes it is useful to look at pairs of dimensions at a time, but the best way is through an interactive 3-dimensional graphing program. In Appendix 3, Figures 18 through 23 show the 3-dimensional solutions, with a pair dimensions represented in each graph. Each of these pictures shows a red circle that represents each sequence in scaled string-distance space, as located by the MDS algorithm. Each line represents the relation between the two MDS extracted dimensions for the sequence, and input (TotalVendorCount, VendorExperience), outcome (SC_to_AA), and automation (AutoPCT) variables.

For example, the entry phase graphs (Figure 18, Figure 19, Figure 20) show that the MDS algorithm located each of the sequences according to their string distance in a coordinate plain. Automation seems to have a nearly vertical slope in all three graphs, indicating that it has a similar relationship in all three dimensional pairs. Total Vendor Count seems to have a similar relation with V1, V3 as with V2, V3, but a different one with V1, V2. We can also see that these lines do not explain much of the variation in these spaces. It would be difficult to imagine a good fit for any drawn line for the entry phase sequences, but the approval phase seems to be linear at least across some of the pairs of dimensions.

What do these Dimensions Mean?

The qualitative examination of sequences and the three dimensional solution indicated some regularities in the way that the algorithm located the sequences, at least at the extreme points on each dimension. Some of the variables, namely the third dimension for the entry and approval might have a relationship that is difficult to discern, or there might be no relationship at all. Other variables, specifically in the approval phase, seemed to be associated with the length of the sequence. In some cases, I was able to observe some patterns that were more common at one end of the dimension than others, but I was still unable to understand what the dimensions mean. The scaled string distance appeared to be identifying differences in the sequences, but the ordering of the sequences themselves was difficult to interpret.

Exploring the relationships between the sequence dimensions and other variables by using multiple regression and visualization suggests several things. First, the amount of explanatory power of these dimensions may be lower than is advised by literature. Some of this may be due to information requirements, an omitted variable, or as indicated in many of the scatter plots, the lack of any linear relationship in the underlying data. Second, automation appears to have the strongest relationship with the extracted dimensions, especially for the Entry phase.

Third, there may be divergence in how these relationships are expressed between the entry and approval phases. There were differences in all parameters that the regressions found with regard to the relationships with variables of interest. The amount of explained variance, the pattern of significance, and most importantly the coefficients of the variables indicate that the relationships between inputs, outcomes, and automation may be different for the entry phase as compared to the approval phase. Next I test the theory presented in chapter three by examining these relationships using path analysis.

Evaluating the Research Questions

In this section, I posit answers to my research questions, exploring the concepts of buffering and the role of automation. I use the three-dimensional MDS solution, but locate the extracted variables in a structural model that is run using partial least squares analysis. This method allows me to include all of the variables of interest, and evaluate each research question as a whole. This technique also allows the creation of constructs to represent the concepts of environmental input, dimensions of the process, outcomes, and automation. I first examine the input-process-outcome model, separately for the entry and approval phases. Then I add automation, and discuss how the use of technology changes the impact and nature of buffering.

This analysis is complicated by the fact that I have two processes or subroutines that take place within the overall invoicing routine at the research site. For each model, I have two sets of data for consideration. This is interesting, because it allows the model to express different relationships between inputs, process, outcomes, and automation for the entry phase and the approval phase. As noted in Figure 8, a MDS projection correctly clustered these subroutines as being different from each other, and the results from the PLS analysis point to differences between their relationships in the model as well.

Construct Formation

In this analysis, I define the input construct as a linear combination of the log of the average invoice amount, the total vendor count, and vendor experience. The process construct is a combination of the three extracted variables relating each sequence in scaled string distance space. The outcome is measured by a single variable, the length of time from scanning to complete approval (SC_to_AA). I have included a covariance and correlation matrix for all of these variables, with the measures for each process in Appendix 3, Table 38 through Table 41.

Because of the way that these constructs are defined for this analysis, they are considered formative. This means that each of the variables may tap into a different facet of the construct, and that each construct may in fact be multidimensional (Petter, Straub, & Rai, 2007). Reflective constructs are different in that each item for a construct is expected to move with the other items for that construct: together they are one-dimensional.

This means that the items that form my constructs need not correlate, and also that they are not measured with error. For example, there was no instrument that measured the amount of the invoice, so there is no way to introduce measurement error. Instead of measurement error, formative constructs have an error term associated with the construct itself that represents misfit of the items with the construct, miss-specified items and items that may be missing from the construct (Diamantopoulos & Winklhofer, 2001; Jarvis, Mackenzie, & Podsakoff, 2003; Peters & Saidin, 2000). The arrows in Figure 11 through

Figure 14 point from the items towards the construct as is customary (Diamantopoulos & Winklhofer, 2001; Jarvis et al., 2003; Peters & Saidin, 2000), and smartPLS deals with them appropriately as formative.

Inputs, Process, Outcome: Buffering

Figure 11 shows the coefficients and R^2 for the input-process-outcome buffering model. The explained variance in the outcome variable is low, and the input and process constructs do not load well from their components. The three internal paths are significant, but overall this is not a great model. I also evaluated splitting the input construct into Invoice Amount and Vendor Experience. This produced a better fit (significant loadings) for the input construct, but the internal paths were insignificant. For comparison to the approval phase, I decided to leave this model as it is presented here.

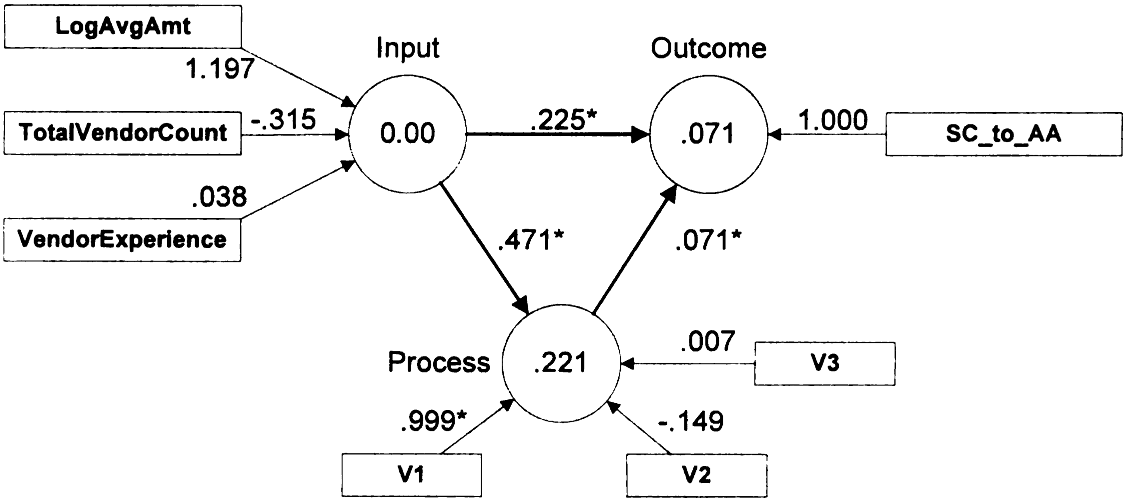


Figure 11: Coefficient and explained variance for input-process-outcome buffering model, entry phase

The approval phase buffering model shows a number of improvements over the entry phase model. First, the amount of explained variance in outcomes is approaching practical significance, albeit at a small level. Interestingly, there is not much variance in the extracted process variables being explained here. The paths between input and outcome, input and process, process and outcome are all significant.

The relative sizes of coefficients suggest a theoretically relevant story. The input-process and process-outcome paths are almost double that of the direct path between input and outcome. Comparing this model to the entry phase model, one can conclude that the entry phase has little connection to outcomes to begin with, but the process is not a core buffering mechanism. In the approval phase, the process appears to be a partial mediator of the relationship between input and outcome, implying that the process buffers the variance in the inputs from impacting the outcomes of the routine.

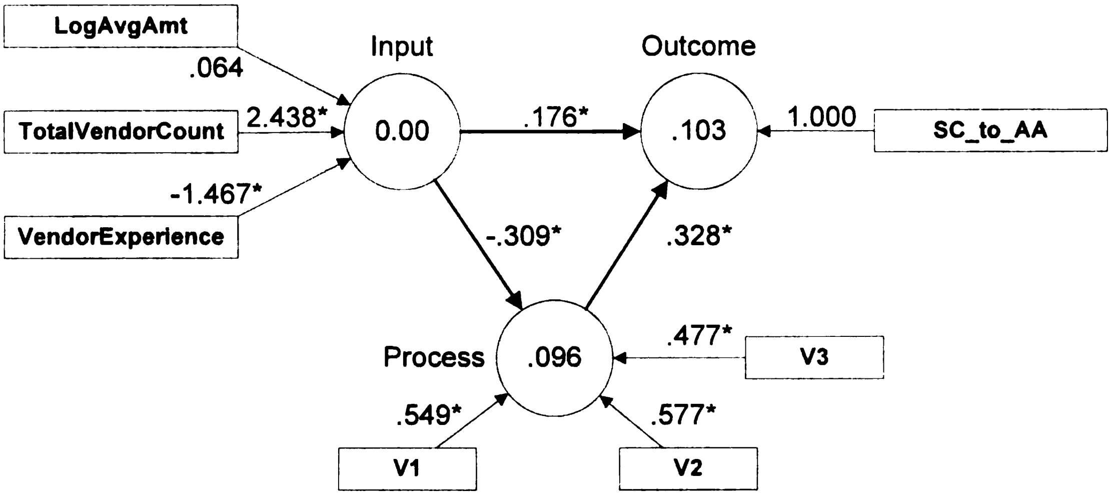


Figure 12: Coefficient and explained variance for input-process-outcome buffering model, approval phase

The Role of Automation and Information Technology Use

Now that the model evaluating how processes can buffer environmental variance has been evaluated, I explore how automation fits into the picture. Figure 13 shows the results of adding the percentage of automation within a process to the entry phase model. Interestingly, the amount of explained variance in outcomes dropped, but this model explains much more of the process. The pattern of construct loadings was superior as compared to the entry model without automation. When examining the coefficients of the paths, we can see how automation may be a buffer within this routine, rather than the process alone.

Inputs drive the amount of automation in a process, and automation strongly drives the process, while the relationships of input-process and input-outcomes are very weak. Given the size of process-outcome and automation-outcome relationships, we can see that the process acts as a buffer most strongly through automation, not via the inputs. This model does support buffering, but it highlights a very complex information system impact through the use of automation features. The entry phase does loosely couple inputs and outcomes, but it is through automation's impact on the process, rather than simple contingent actions acting as the mechanism. This indicates that buffering mechanisms can use a combination of actions and technology, rather than solely on the expression of actions.

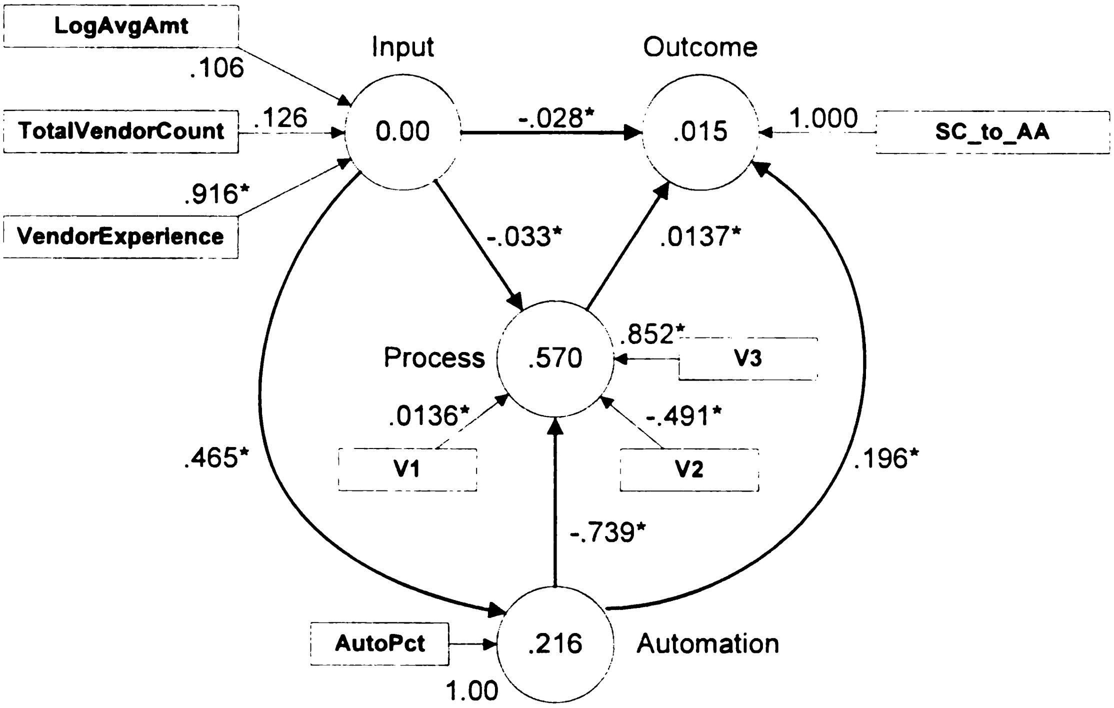


Figure 13: Coefficient and explained variance for input-process-outcome with automation model, entry phase

The approval phase held additional surprises and implications from the addition of automation. The outcome variance explained (R^2) improved, indicating that automation does add more information to the model. On the other hand, this model explains less of the process variance. All paths were significant; construct loadings as well as internal paths, indicating that this is may be a sufficiently acceptable model for this data structure. This model is consistent with the approval buffering model without automation, but in this phase of the invoicing routine, we find that inputs do affect processes directly *and* through automation. Looking at the relative sizes of the path coefficients, we see that the process emerges as a stronger buffer, in that the input-outcome connection is less than half the size of the input-process and process-outcome relationships.

Interestingly, the size of the automation-outcome coefficient is smaller than the input-outcome coefficient, but the connection between input and automation remains high. This implies that the process acts as a buffer, but automation also has some buffering content in how it affects the process. If one were simply studying the impact of automation on outcomes, the results would completely miss its impact through the process as a mediator. This highlights how innovative the methods are in this study, and represents a new class of IT impacts that have not yet been investigated or theorized.

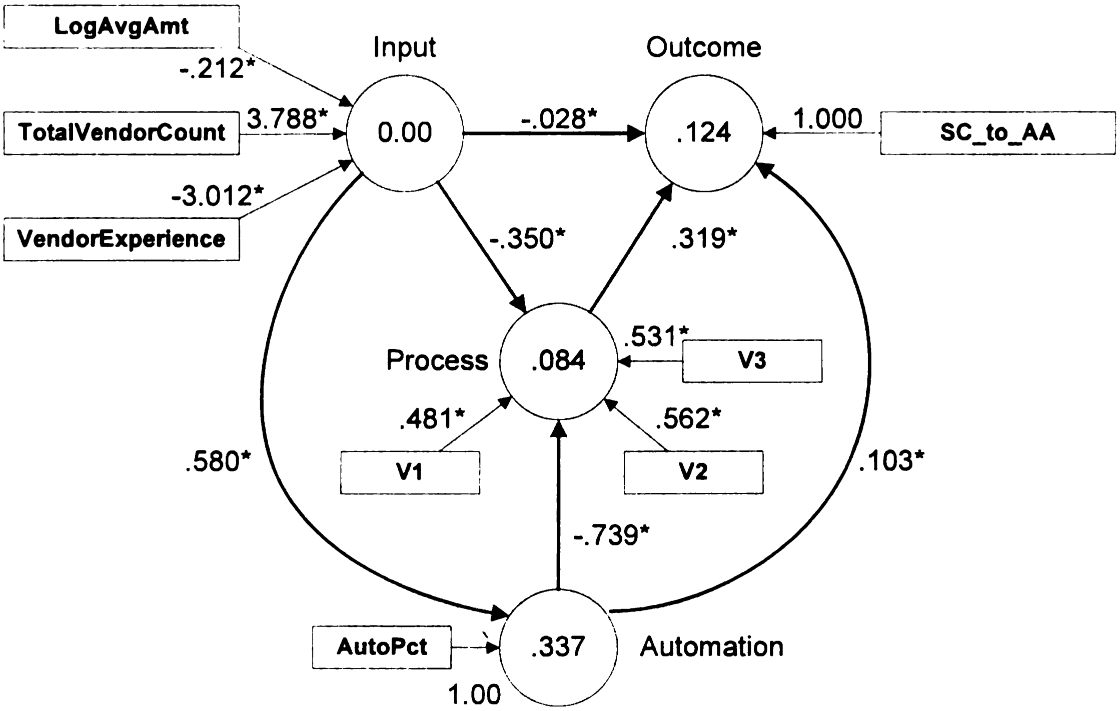


Figure 14: Coefficient and explained variance for input-process-outcome with automation model, approval phase

Comparing the effects of automation on the entry model to that obtained through the approval model provide additional insights. It appears that the role of automation is markedly different between the two phases of the invoice process. In the entry phase, the process does mediate the variance between the inputs and outcomes, but only through the mechanisms of automation. In the approval phase, automation seems to be an adjunct or secondary buffer, as the process does buffer independently of automation.

This indicates some of the complexities of understanding the impacts of information technology use. In general, the entry phase has a higher mean automation, but the distribution in the approval phase is much wider. Even though I have only one measure of automation that is calculated the same between the subroutines, there is heterogeneity in what automation *means* for each process. The task context is different, but also the nature of use may be different between the two phases. For example, many of the actions that are automated in the entry phase are data entry and forwarding. In the approval phase, much of the automated tasks are notifications and forwarding for further approval.

Discussion

The number of choices given to the researcher with this set of methods results in a complex picture to interpret. While the dimensions extracted from the process were difficult to interpret with the available variables through regression and visualization, the path analysis was consistent with the Markov results. The low amount of explained variance may be due to the information content in the sequences compared to the total information available based on the lexicon and sequence length.

In 2000 sequences I could have at most 2000 different sequences, but there were only 241 distinctly different ones. This means that the sample of sequences only has a fraction of the information contained in the whole set of possibilities. This smaller set of information is the only basis that the MDS algorithm can use to find differences and variance between the sequences. Normal calculations of r-square use the explained variance divided by the total variance. I believe that the total variance I have in my case is smaller than normally found, leading to a smaller SST and r-square being unnaturally low. This is may be correctable statistically but is beyond the scope of my study.

Evaluating model quality is difficult in models including formative constructs, because measures of internal consistency are only useful concepts when applied to reflective constructs (Diamantopoulos & Winklhofer, 2001; Jarvis et al., 2003; Peters & Saidin, 2000). Typically, one must use either a correlation to connecting reflective scales or consider the relative sizes of coefficients and explained variance to ascertain the value and quality of a given model. Given the low explained variance, and difficulty with significance in some paths, it would be helpful for validation through other regression techniques in addition to PLS. When my model was processed using seemingly unrelated regression (SUR), similar results with significance and explained variance were obtained. When my model was evaluated using three-stage least squares (3SLS), the model failed because of endogeneity and identification issues.

The fact that all of the constructs in my model are formative has implications on the quality of the model. This means that I have an identification problem, because there is some indeterminacy between the error terms and the scale of measurement (Jarvis et al., 2003). Possible solutions include setting one of the indicator paths to 1 or adding a reflective construct to the model. The best time to solve this identification problem is at the research design stage, before analysis has begun (Diamantopoulos & Winklhofer, 2001; Petter et al., 2007). Since I do not have the option of adding a new reflective construct, my options may be limited with the current analysis, but a respecification of the model or the use of a different regression technique may prove useful.

There are significant issues with non-linearity and non-normality in the data. The ultimate DV is not continuous; rather it is a count of the number of days it takes to process an invoice. The nature of this variable may indicate that probit or logit analysis with different distributional assumptions may be appropriate. The MDS dimensions may exacerbate this issue, given that there are only 206 different values being applied to 2000 entry sequences.

An additional reason the explained variance is low is the likely omission of other variables that may explain these relationships better. For example, I was unable to categorize the vendors or their products, due to them being written in Norwegian. There may also be organizationally relevant variables that the workflow system does not capture. Despite these challenges, I believe that these analyses lead to some important findings, and implications that can be generalized.

The results have implications for the concept of buffering. Given the differences between the entry and approval phases in how the process protects outcomes from environmental variety, this indicates that we may need to think a little differently about how contingent actions act as a buffering mechanism within business processes. Rather than seeing the business process as a whole, I find that buffering occurs differently in subprocesses of the invoicing routine. This suggests that the environment may impinge upon different *sections* of a routine, and create points of buffering sections within a given sequence. Routines may have internal heterogeneity in how variety in a routine is harnessed to buffer environmental variety. We may have to look deeper within the set of sequences a routine generates to find sections that are buffers within the routine itself, rather than looking at a whole routine as a buffering mechanism.

The results also indicate some changes in how we view the concept of information system use. We already know that there is heterogeneity in how individuals and organizations use information technology, and this challenges our ability to study the impact of IT in a generalizable way. This study shows that there is heterogeneity also found *within* an organization as to how different subprocesses supported by information technology have differential impacts of the use of IT, adding a layer of complexity to studies of IT impact that have not previously been examined. For example, one impact of IT on the entry process is that of a buffering mechanism that substitutes for contingent actions. The approval process is a complement to contingent actions as a buffer. In this way, information technology may impart a different class of impact from what has been previously theorized and empirically tested.

The implication is that we can achieve different results from the application of technology in different subsections of a business process. If technology use has a negative impact on the front half of a process, but a positive impact on the back half of a process, a study that looked at the immediate consequences may find no relationship, when in fact there were two effects that cancel each other out. This highlights how complex the relationship between IT use, processes, and outcomes may actually be, and also some of the difficulties that researchers and managers have with evaluating the impact of IT investment, adoption, assimilation, and even use.

Chapter 7—Discussion and Limitations

Introduction

In this chapter I integrate the results with theory, and discuss implications of the study on literature and practice. I then describe some limitations of the approach I undertook, and close with some directions for future research, including a new research design. Overall, this research is designed to be explanatory. I begin with theory, a set of a priori assumptions about the world and adopt a confirmatory approach. The methods I used to evaluate the research questions within my theoretic framework indicate a more exploratory approach. While these methods have been applied in other areas relating to processes and sequences, their application to workflow data to test theory has been absent. There is a tension between exploration and explanation in this work that will be resolved with future work and further study.

Theoretical Implications

In this section I connect the results of my work with the larger conversations taking place in the literatures of organization theory, routines, business process management, and information technology impact. I also explore extensions and improvements to this research in the various areas that help us understand buffering, process management, and the impact of information technology.

Organizational Theory

Organizational theorists have been interested in how the environment affects organizational systems as soon as they perceived its open nature (Scott & Davis, 2007). While there have been several empirical studies of Thompson's (1967) theory of buffering (Cooper & Smith, 1992; Koberg, 1988; Sorenson, 2003) and recent theoretical developments (Lynn, 2005; Yan & Louis, 1999), there have not been any studies of buffering at the business process level. More importantly, none has embraced the perspective that examines the actual actions that take place as a buffering mechanism. The results of this dissertation confirm the buffering of environmental variety: Outcomes are weakly related to inputs. Interestingly, inputs and automation drive changes in the process that are transmitted to the outcomes.

This dissertation serves as an exemplar of applying the theory of organizational routines to improve our understanding of organizational actions and structure. Empirical studies of organizational routines are rare mainly because of the difficulty and cost of obtaining and analyzing data tracking actual events relating to hundreds of process executions (Pentland et al., 2009d). The work presented here significantly adds to the literature on organizational routines in at least three ways.

First, the Markov approach gives a measure of the probabilistic relationship between temporally connected actions in the routine. This connects to the conceptualization of organizational routines as habits or 'dispositions' (Schulz, 2008), and represents one of the methods we suggest to empirically compare routines (Pentland et al., 2009a). Second, measures of sequential variety indicate the amount of variation in the choice and order of actions within a sequence. If we look at this attribute of an organizational routine over time, we can explore aspects of endogenous change within a routine (Pentland et al., 2009b, 2009c), but also the effects of managerial intervention on a process.

Third, organizational learning can have an equivocal effect on the variety in a process. On one hand, as systems learn which sequences do not work or are undesirable, these sequences are pruned from the set of possibilities, leading to a reduction in sequential variety. Conversely, the act of learning new ways of performing an organizational routine would involve trials of candidate sequences, leading to an increase of sequential variety. Sequential variety may represent the natural 'repertoire' of the different ways an organizational routine can be performed under various stimuli or conditions in addition to improvisation or errors.

Business Process Management

This work rejects the black-box approach to understanding and managing business processes. While a few scholars in the field are beginning to understand the implications this change in perspective (Melão & Pidd, 2000, 2008), it is definitely not widespread. By understanding the actions that take place in-situ, and studying how people and technology interact, scholars of business process management can connect to other related literatures such as organizational routines and management.

The synthesis of flexibility and stability represents an extension of the BPR/BPM literatures, and can be found in areas such as lean and custom manufacturing, services, and high-reliability organizations. Despite the rise of these innovative strategies, typical literature in the management of business processes often begins with a perspective of conformance and matching process executions to documented standards (Singh et al., 2009). This dissertation begins with a different perspective: embracing variety in execution to understand its antecedents and consequences. In this way, I seek to be one of the bridges between the organizational routines literature and that relating to business process management.

Another common feature of business process management research is the use of 'typical' rather than 'actual' representations of the process (Singh et al., 2009). This means a focus on the abstract features of the usual process, or what steps *should* be performed within the process, usually obtained through interviews. Research that uses 'actual' representations uses observational data in some way to discern what actions are expressed within the business process. A highly prolific group of the BPM scholars proposes the use of workflow mining to automatically extract and visualize patterns of a process based on the action logs that are recorded by the workflow software (Agrawal et al., 1998; Agrawal & Srikant, 1995; van der Aalst, Desel, & Oberweis, 2000; van der Aalst, ter Hofstede, & Dumas, 2005; van der Aalst & van Dongen, 2002; van der Aalst et al., 2003; van der Aalst & Weijters, 2004; van der Aalst et al., 2004). Van der Aalst and his colleagues suggest the use of workflow mining in the investigation of organizationally relevant research questions in addition to conformance and pattern extraction (van der Aalst et al., 2003; van der Aalst & Weijters, 2004) and this dissertation answers their call.

The results of this dissertation point to the management of variety through contingently expressed action as a method of protecting the core. While there has been a widespread understanding of the systemic properties of buffering, there have not been any examples of an empirical test that can be applied directly to a business process. The sequential variety analysis supports sequential variety as an expression of contingently expressed actions. Queuing models and other management science techniques represent one method analyzing and designing business processes for buffering. This dissertation demonstrates a different view of incorporating specific *actions* as the central feature of business process management and technology as a core feature of digitally enabled routines.

IT Impact

Studying the immediate antecedents and consequences of technology situated in a single business process allows the isolation of specific use effects. This explores the moderation effect of the use construct, and complements the firm-level, organizational, and behavioral impact literatures. The impact of IT can move beyond a study of investment (Brynjolfsson & Hitt, 1995) or adoption (Venkatesh, Morris, Davis, & Davis, 2003), into research questions that relate to exactly *how* IT drives value. This can occur by studying the enablement and constraint of organizational actions as a primary impact of information technology.

The results of this dissertation point to automation as a discriminator among patterns of action. From the Markov results, heterogeneity of the process was found among groups of sequences that varied with the amount of automation expressed in the performance. The sequential variety analysis confirmed this result and revealed automation as a strong player in the buffering of environmental variety, beyond its impact on the outcomes of the process. This was a surprising result, and points to a new finding from the substitution of IT use for labor—buffering. We have known for some time that IT is a substitute for other forms of input such as labor and ordinary capital (Dewan & Min, 1997), but less was known about *how* it can substitute.

What is interesting here is that automational technologies are typically seen as a substitute for human labor, when decision-making needs can be anticipated and relevant stimuli identified. From the cybernetic world-view, a system exhibits a variety of responses that is equal to the variety in the inputs. When tasks are automated, the decision-making as to which response is appropriate is designed into the system, such that the match between stimuli and response is predetermined, and that the set of stimuli and responses can be developed before the set of rules are ingrained into the system. In most cases the set of automated responses to stimuli is much smaller than would be possible if decisions were guided intelligently at the moment of execution. What I am proposing from the findings in this dissertation is that automational technologies, despite the reduction of flexibility as compared to a manual system, can still act to buffer a process from variety in inputs. This can occur through several ways.

First, the tasks that are automated can be general purpose, where a given response can respond appropriately to many different kinds of stimuli. This was observed in the mail sorting example, where machines to sort the mail using OCR did not discriminate between Helvetica, Arial, or Times Roman fonts, but the OCR applied equally to each. Second, given the ability of the system to correctly discriminate between types of stimuli, automation may allow a more consistent application of rules and lead to a more easily manageable organizational system. These two features of automational aspects of an organizational information system show how automation can be used to buffer and protect the technical core of an organization. Taken together with the earlier observation of the ability within sections of an organizational routine to act as a buffer, the impact of automational IT can also be seen to act as a substitute buffer to process-based buffering.

Future research could explore other aspects of information technology impact such as how the features of a given system support or improve informating up, down or sideways. It may be possible to examine the actions taking place within the current dataset to discern a typology of different actions that could be theoretically interesting. In general terms, I see actions that are in the following categories: information processing, decision-making, and coordination. There may be other, more theoretically driven categorization schemes.

Methodological Implications

The methods used in this dissertation can be used to understand organizational behavior phenomena in many other areas. I use them as an attempt to synthesize a middle path between qualitative and quantitative research of organizations. While workflow mining doesn't provide the richness and depth of understanding of causality in organizational processes, it does allow the development of statistical conclusions that focus on what really happens rather than mathematical relationships between numerical proxies for actions. One way to describe this approach might be the "variance of processes" or "process-oriented variance analysis".

In most cases, research is either process-based or variance-based (Markus & Robey, 1988). There have been some proponents of studying the properties of processes (Monge, 1990) and also those who suggest different ways to study processes themselves (Langley, 1999; Sabherwal & Robey, 1995; van de Ven, Angle, & Poole, 1989; van de Ven & Poole, 1990). It appears that these are competing perspectives, studying similar phenomena from different directions. I view them as complementary and not mutually exclusive within the same research plan. By associating different patterns of action (representations of the process) with the variance of inputs and outputs, I am integrating the quantitative strategy outlined by Langley (1999, p. 697) with the evaluation of properties of processes over time proposed by Monge (1990).

Practical Implications

In this section, I discuss how the results of this dissertation can help the practice of management, information systems design, and information systems use. I also explore implications for the education of managers and IS professionals.

Managerial Impacts

Managers, especially those of boundary business processes, must understand the complex interaction of environment, process, and outcomes. Their ability to manage uncertainty is challenged by the need for stability and control over the process. As they seek creative ways to simultaneously improve quality and efficiency, a focus on the specific causes of expressed patterns of action represents a different way to look at managing processes than is currently taught in business schools today.

Focused on the black-box approach to managing processes, programs such as TQM and six-sigma measure and statistically measure the outcomes from a process, with an emphasis on control. Process standards such as ISO 9000 treat processes as fixed, and deviation from documentation is considered a sign of poor process execution. As business schools (and resulting managers) follow these programs, they forgo the opportunity to dynamically monitor and adjust the processes themselves both in advance and at the time of execution. The worldview described in this dissertation represents an opportunity for managers to shift their thinking to new paradigms of managing processes. This has an impact beyond traditional management perspectives and can be applied to supply chain, remanufacturing, and service provision among other areas of managerial practice.

Without an understanding of the drivers and consequences of specific patterns being expressed within processes, managers must continue to use the tools of statistical process control such as TQM and six-sigma to achieve some measure of regulation. The discovery of these drivers and outcomes within a business process represent a new mode of management that was previously unavailable to be implemented. In addition, in areas where statistical process control regimes are less useful such as service provision, managing the sequence and choice of actions within the process holds special promise to give new tools to the practice of management.

Also, this focus on the expression of specific actions increases the ability of managers to discover and learn from their processes. In a world of information overload, discerning patterns and their antecedents and consequence allows the manager to better make sense of the organizational system. Organizational and individual learning can be bolstered by the greater understanding and retention of process-based knowledge that is typically tacit or hidden in the spatially and temporally diffuse business processes that are typically executed in modern organizations.

Finally, managers can now more fully realize the benefits of continuous auditing and assurance (Vasarhelyi & Halper, 1989). This requires at a minimum a good set of IT controls, some form of real-time or near real-time monitoring capability, and the ability of timely release of reports detailing the impact and performance of an assurance and control regime (ISACA Standards Board, 2002). Much of this information becomes available to managers through the use of workflow mining, related technologies as well as managerial intervention (Alles, Brennan, Kogan, & Vasarhelyi, 2006).

Impacts on Information System Design

Designers of information systems need to better understand the consequences of their decisions. Given the tradeoffs between flexibility and control in the designed interactions of users, and the natural tendency of users to innovate and utilize tools for unforeseen proposes, the design of information systems is difficult. The perspective in this dissertation, namely that of studying the actual paths of user behavior within the system, allows IS designers to develop more flexible use-cases, and achieve synergy between control and elasticity of the IS-enabled process.

There has been some research related to the design of web sites involving the collection and interpretation of 'clickstream' data mined from logs of web servers (Kosala & Blockeel, 2000). Typically, researchers have focused on discovering patterns of user interaction to categorize users (Buchner, Baumgarten, Anand, Mulvenna, & Hughes, 1999; Cooley, 2000; Cooley, Mobasher, & Srivastava, 1999), and improve the user experience (El-Ramly, Stroulia, & Sorenson, 2002). In this literature, there has been less interest in the management of the web usage process, but rather in the practical aspects of design and development of usable systems. The research presented in this dissertation represents a complimentary view to the models of web usage, as a process that the user and their characteristics become inputs, and the outcomes can be measured in terms of success, failure, effort expended or satisfaction.

This dissertation explores automation as the core feature within workflow systems. Given that there were differential effects of technology use on the process, information system designers may need to look closer to sections or subprocesses within a business process for appropriate system designs. For example, processes should have automation and control where appropriate, yet have flexibility where it is needed within sections of the process. There may be reasons to implement controls in the system to reflect physical constraints, business rules, institutional and social norms, but these should only impact the business process during specific times, places or within action sequences that are expressed.

In general, this research supports the following principles of organizational and information system design:

- Automate where information needs are sufficient to determine the appropriate actions without human decision-making (Ashby, 1958, 1968; Cyert & March, 1963).
- Make information available to support human decision-making and conserve the scarce resource of attention (Simon, 1973).
- Coordinate between individuals when resources (especially knowledge or information) are interdependent (Crowston, 1997; Grant, 1996; Malone et al., 1999).

Workflow and other organizational technologies can be designed for monitoring and control over a business process. If this perspective is followed too far, the reduction of flexibility may cost more than the benefits that are enabled through the use of the system. This understanding should be core to the design of information systems, especially ones with organization-wide effects such as ERP and workflow systems. This is not a new insight (Merton, 1936), and there are some scholars that see ERP as the new 'iron cage' (Gosain, 2004), but these ideas have not become widespread in ISD education.

Impact on Information System Users

Similarly, users of information systems must understand what they give up in terms of flexibility when they adopt a particular information systems solution. While some vendors are starting to add exception handling and more flexibility to workflow, the organizational costs of too much control over a process are less understood. There is a suggestion here that standards such as ISO 9000 may have hidden costs beyond documentation and certification through the restriction of flexibility in organizational action.

Limitations

The main limitation to this study is the difficulty in interpreting and integrating the analyses. The multiple analyses were qualitatively consistent, and yet highlighted different perspectives of the performance of routines. While using contingency table tests to compare the transition matrices is an effective way to explore the group membership of various sets of sequences, the results of these tests give binary responses. There is no measure of *how* different the sequences are from each other. From the scaled string distance approach, there is a much better measure of sequence distance, but the extracted dimensions are difficult to interpret. There is no measure of what the differences *mean*. Additional analyses or extensions of these methods that can integrate perspectives and give a more complete picture of the antecedents and consequences of sequential variety must be performed across a variety of contexts.

Another limitation relates to the source and characteristics of the data. There may be actions that are part of the invoicing entry and approval subroutines that occur outside the purview of the workflow system. As with any observation of organizational actions, research design choices, politics, cognitive limits to inspection, and many other factors determine what is available for analysis by researchers. This does not invalidate the findings of any study, but may limit the types of inferences and conclusions that are possible to be made from such analysis. I believe that additional data and analysis would complement my findings, not contradict them.

I make no claims to understanding the intentions and feelings of participants, the socio-political structure of the organization, or many other aspects of organizational routines that have been theorized or shown to exist. I can make no use of the ostensive aspects of organizational routines at my research site—but this does not affect my ability to answer the research questions. This study focuses on the actual actions as recorded by a workflow system—focusing on the technologically feasible and practically available data for large-scale statistical analysis. I recognize that much of the rich detail that is the hallmark of many studies of organizational routines such as those by Barley (1986; 1990) and Pentland (1992; 1999; Pentland & Rueter, 1994) is not present, but this research represents a complementary rather than contradictory perspective.

Future Research

Given the theoretic, methodological, and practical impacts of this dissertation, there are several natural paths to future research. Some of these could be completed by utilizing the same or similar data, but there are implications beyond organizational theory, business process management, and IT impact. The general form of the Input-Process-Outcome model is easily applied to a number of areas. I realize now that I have developed this worldview and applied it in previous research searching for disturbances in the software development routines in open source software. It has extensions beyond the management of processes, and could be used to study organizational behavior, psychology, accounting (auditing), supply chain, even non-business fields like biology.

Any field that takes a systemic view, and holds some process at the center of inquiry can utilize this basic model.

I would be interested to extend the methods used in this dissertation to analyze different business processes. I should be able to achieve the best connection between inputs, processes and outcomes in some specific, targeted contexts within organizations. Reverse supply chain analysis is the study of how businesses accept returns from customers, and has received recent scholarly attention. I could investigate how the characteristics of the customer and the product would drive how the business process would handle each returned item.

Similarly, a remanufacturing business process would exhibit a variety of actions depending on the qualities and characteristics of the input. Finally, the technical support function may also change its process in response to the joint characteristics of the problem, attitude of the customer, and training of the technician. To the extent that these are digitally enabled through a technology that allows automatic logging and data collection, they may be most appropriate to study buffering and the impact of technology and add to my findings.

Studies could be conducted to better understand the connection of specific inputs to specific patterns of action. One method that has been suggested is related to the Markov approach, but utilizes order statistics (David & Nagaraja, 2004; Rényi, 1953). This approach would model the most probable path through the actions to obtain a 'primal' routine or set of routines. The most probable initial transition becomes the start of the chain, then the most probable transition given that particular starting point. In this manner, a path through the transitions is drawn, based on the probabilities of each one.

Then, the inputs associated with those performances that match exactly can be studied, and the distance of other performances can be computed from the primal routine. This extraction of a primal or modal sequence can be used to increase the power of the scaled sequential variety approach in visualizing the distribution of the routine around this centroid. Also, the use of order statistics integrated with string distance and multidimensional scaling may allow the development of a method with the discriminating power of the Markov approach and the visualization and interpretation potential of the sequential variety approach.

There is more information within the workflow log related to the inputs that was not utilized in this dissertation. I have used the vendors simply as a vehicle of experience, but the nature or line of business for these vendors could be discerned and associated with the processes. Also, I have information about the detail lines on the invoice, such as the number and type of goods that were ordered. These, like the vendor name, are in Norwegian, and would necessitate the use of a native speaker to translate them, and they would then need to be categorized and coded. The use of semantic models might be able to be used if translation is not an option, not focusing on the meaning, but the connection of symbols to processes.

Another extension using the same data (and similar methods) would be to explore the impact and interaction of the action network with the social network that completed the work. In some ways, this would be just be adding a mode of connection between social actors for every action. Both methods used in this dissertation could be applied to this data. Three Markov matrices could be extracted: the action-action transitions, social-social transitions and the action-social (role) transitions. The string-distance and MDS approach could be applied to the sequences of people, and also to sequences of people-actions. Candidate research questions are easy to visualize. What has a stronger effect on variation within the process? Is variation driven by changing actions, changing people, or changing roles over time or various combinations thereof?

Finally, learning effects can be examined. Since I have data from the initial installation of the software, I could examine how patterns of action change over time in relation to efficiency. This examination of the learning curve could consider the relationship between sequential variety and efficiency over time. The intuition is that with experience, people tend to try things, and learn what not to do, leading to a drop in sequential variety with experience. Interestingly, sequential variety seems to be increasing over time, meaning that the repertoire of organizational routines may be increasing with experience. This highlights the importance of understanding the impact of sequential variety on learning and learning models both in theory and practice.

Another extension of this work into organizational learning allows a much more micro focus on how individual 'learnings' are combined to form the traditional logarithmic form of the learning curve. Because learning occurs at several levels of analysis, from individual, to between individual to group and organization, it would be interesting to map out what the experience curves are at each level, and how they interact between levels to allow the organization to learn from its environment and prosper. Also, economies of scope in learning can be explored, moving beyond the experience curve (economies of scale) and examine the transference and retention of different types of knowledge within the organization.

I can envision an application of March's (1991) learning model to the data. Docking his environment-organization-outcome simulation model to the data in this dissertation could be attempted. This would represent a contingent-fit approach to learning, and would connect to related recent research such as that by (Miller, Zhao, & Calantone, 2006). I can also envision the analysis of data obtained through experiments similar to Cohen and Bacdayan (1994).

The methods I have utilized are not limited to creative uses with the current set of data, as they can be applied to many various areas. Given the rise of organization-wide information systems such as ERP, this may make much more process data available. If the correct site could be found, I would like to apply these methods to the entire organizational system, as the different business processes interact. This type of analysis would be complex, and probably beyond the ability of personal computing technology to implement, but it would allow the investigation of many interesting research questions.

Appendix 1: String Matching Distance

adapted from (Pentland, 2003b; Pentland et al., 2009d)

This measure is defined as the average distance between each pair of observed sequences. A standard technique for measuring the distance between two sequences that may vary in length is called optimal string matching (Sankoff & Kruskal, 1983; Abbott & Hrycak, 1990; Gribskov & Devereux 1992; Sabherwal & Robey, 1993). String matching has been used extensively in molecular biology to compare protein sequences, such as DNA. Abbott (1995) provides a review of applications in the social sciences.

The distance between two strings can be computed by counting up the number of operations needed to transform one string into the other. The operations include substituting one element for another, or inserting or deleting elements. Each operation has a cost, and the distance between the strings is the total cost. In this paper, all of these costs were set equal to one, but could be adjusted to account for similarity of actions, as discussed below. The technique is called 'optimal' string matching because it finds the lowest cost set of operations to accomplish the transformation, thus insuring that the computed distances are unique and well-behaved (e.g., they obey the triangle inequality: d(A,B) + d(B,C) >= d(A,C)). Distances computed in this way are called Levenshtein distances (Sankoff & Kruskal, 1983).

Observations can be represented in an N x M array of events, where each row corresponds to one iteration of the process, as seen in equation 3:

$$\text{Observed sequences} = \text{S} = \begin{bmatrix} e_{11} & e_{12} & e_{13} & . & . & . \\ e_{21} & e_{22} & e_{23} & e_{24} & \cdots & e_{2M} \\ \cdots & \cdots & \cdots & \cdots & . & . \\ e_{N1} & e_{N2} & e_{N3} & e_{N4} & \cdots & e_{NM} \end{bmatrix} \quad (3)$$

where N = the number of observed sequences and M = number of events in the longest sequence. Since the length of the observed sequences may vary, this array can have a 'ragged' edge (signified in equation 3 by '.'). This representation includes each observation in its entirety.

To estimate the variation in a set of sequences like those in Equation 3, we can compute the distance between each sequence and every other sequence. If the sequences were all identical, then the distances would all be equal to zero. If the sequences diverged from each other in a single element (e.g., 'aaa', 'aba'), then the distances would all be equal to one. As the differences between the sequences become more pronounced, the distances increase. Thus, a convenient and meaningful measure of variety in a set of sequences is simply the average of distances between all pairs of observations, shown in Equation 4:

$$\textbf{Average distance} = \frac{1}{n(n-1)/2} \sum_{i=1}^{N} \sum_{j=i}^{N} d(i,j) \quad (4)$$

where N equals the number of observed sequences and d(i,j) equals the Levenshtein distance between each sequence. The factor n(n-1)/2 is simply the number of pairs in a set of n sequences. Alternatively, the entire matrix of relative distances between sequences can be used as in this dissertation.

Appendix 2: Markov Analysis

Action	1	2	3	4	5	6	7	8	9	10	12	20	23	25	27	28
1	470	0	61	0	0	0	0	2	49	11	436	0	23	155	1	0
2	70	0	0	0	0	0	0	0	1	8	0	0	0	0	0	0
3	0	65	0	120	10	0	1	3	0	0	517	156	275	157	1	0
4	0	0	337	0	38	3	3	154	0	4	3	319	413	0	0	0
5	0	0	19	208	0	28	18	221	0	0	0	39	0	0	0	0
6	1	0	322	166	1	0	0	0	0	18	18	7	0	0	0	0
7	1	0	59	10	0	448	0	0	1	6	5	2	0	0	0	0
8	0	0	12	29	0	46	439	0	1	3	3	0	0	0	0	0
9	7	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0
10	468	38	7	761	1	1	0	3	8	0	9	0	1	0	0	0
12	242	82	1	0	0	1	11	9	1	457	0	10	107	245	4	1
20	4	0	110	0	0	6	60	104	0	36	213	0	0	0	0	0
23	15	151	77	0	0	0	0	0	0	0	2	0	1080	474	53	3
25	247	379	222	0	0	0	0	0	0	7	73	0	7	146	1	4
27	4	4	4	0	0	0	0	0	0	0	0	0	2	35	34	4
28	0	3	1	0	0	0	0	0	0	0	0	0	5	2	0	0

Table 32: Transition matrix for entry phase

Quartogram (65536)	n_4	$n_i \log n_i$	Trigram (4096)	n_3	$n_i \log n_i$	Digram (256)	n_2	$n_i \log n_i$	Symbol (16)	n_1	$n_i \log n_i$
Sum	14346	105661.6	Sum	16346	135661.2	Sum	18346	173253.8	Sum	20346	221817
$\hat{H}_{quartogram} = \log n - \frac{1}{n}(\sum_{i=1}^{65536} n_i \log_2 n_i)$			$\hat{H}_{trigram} = \log n - \frac{1}{n}(\sum_{i=1}^{4096} n_i \log_2 n_i)$			$\hat{H}_{digram} = \log n - \frac{1}{n}(\sum_{i=1}^{256} n_i \log_2 n_i)$			$\hat{H}_{symbol} = \log n - \frac{1}{n}\sum_{i=1}^{c} n_i \log_2 n_i$		
$\hat{H}_{qua} = \log 14346 - 14346^{-1}(105661.6)$ $= 13.8084 - 7.3652$ $= 6.4431$			$\hat{H}_{tri} = \log 16346 - 16346^{-1}(135661.2)$ $= 13.9967 - 8.2994$ $= 5.6973$			$\hat{H}_{di} = \log 18346 - 16346^{-1}(135661.2)$ $= 14.1632 - 8.2994$ $= 4.7195$			$\hat{H}_{sym} = \log 20346 - 20346^{-1}$ (221817.7) $= 3.4102$		
$\hat{H}_4 = \hat{H}_{quartogram} - \hat{H}_{trigram}$ $= 6.4431 - 5.6973$ $= 0.7458$			$\hat{H}_3 = \hat{H}_{trigram} - \hat{H}_{digram}$ $= 5.6973 - 4.7195$ $= 0.9778$			$\hat{H}_2 = \hat{H}_{digram} - \hat{H}_{symbol}$ $= 4.7195 - 4.7195$ $= 1.3093$			$\hat{H}_1 = \hat{H}_{symbol}$ $\hat{H}_0 = \log c = 4$		
$\hat{T}_3 = \hat{H}_3 - \hat{H}_4$ $= 0.9778 - 0.7458$ $= 0.2320$			$\hat{T}_2 = \hat{H}_2 - \hat{H}_3$ $= 1.3093 - 0.9778$ $= 0.3315$			$\hat{T}_1 = \hat{H}_1 - \hat{H}_2$ $= 3.4102 - 1.3093$ $= 2.1009$			$\hat{T}_0 = \hat{H}_0 - \hat{H}_1$ $= 4 - 3.4102$ $= 0.5898$		
$X^2 = 1.3863 n_4 T_3$ $X^2 = 1.3863\ (14346)\ (0.2320)$ $X^2 = 4613.4$			$X^2 = 1.3863 n_3 T_2$ $X^2 = 1.3863\ (16346)\ (0.3315)$ $X^2 = 7512.3$			$X^2 = 1.3863 n_2 T_1$ $X^2 = 1.3863\ (18346)\ (2.1009)$ $X^2 = 53431.5$					
$df = c^{n-1}(c-1)^2$ $df = 256(15)^2$ $df = 57600$			$df = c^{n-1}(c-1)^2$ $df = 16(15)^2$ $df = 3600$			$df = c^{n-1}(c-1)^2$ $df = 1(15)^2$ $df = 225$					
p(X^2 =4613.4, df= 57600)=1			p(X^2 =7512.3, df=3600)=2.7534E-277			p(X^2 =53431.5, df=225)=0					

Table 33: Entry phase—Order of the process

Quartogram (160000)	n_i	$n_i \log n_i$	Trigram (8000)	n_i	$n_i \log n_i$	Digram (400)	n_i	$n_i \log n_i$	Symbol (20)	n_i	$n_i \log n_i$
Sum	29826	224160.1	Sum	32668	283888	Sum	35520	362694.4	Sum	38372	455666.9
$\hat{H}_{quartogram} = \log n - \frac{1}{n}(\sum_{i=1}^{160000} n_i \log_2 n_i$			$\hat{H}_{trigram} = \log n - \frac{1}{n}(\sum_{i=1}^{8000} n_i \log_2 n_i)$			$\hat{H}_{digram} = \log n - \frac{1}{n}(\sum_{i=1}^{400} n_i \log_2 n_i)$			$\hat{H}_{symbol} = \log n - \frac{1}{n}\sum_{i=1}^{c} n_i \log_2 n_i$		
$\hat{H}_{quart.} = \log 29826 - \frac{1}{29826}(224160.1)$ =14.86428-7.515592 =7.34869			$\hat{H}_{tri} = \log 32668 - \frac{1}{32668}(283888)$ =14.99559-8.690094 =6.305497			$\hat{H}_{di} = \log 36115 - \frac{1}{36115}(362694.4)$ = 15.14031-10.04276 = 5.097546			$\hat{H}_{sym} = \log 35520 - \frac{1}{35520}416489$ 15.22777-11.87498 = 3.352782		
$\hat{H}_4 = \hat{H}_{quartogram} - \hat{H}_{trigram}$ = 7.34869 - 6.305497 = 1.043193			$\hat{H}_3 = \hat{H}_{trigram} - \hat{H}_{digram}$ = 6.305497- 5.097546 = 1.207951			$\hat{H}_2 = \hat{H}_{digram} - \hat{H}_{symbol}$ = 5.097546-3.390866 = 1.744764			$\hat{H}_1 = \hat{H}_{symbol}$ $\hat{H}_0 = \log c = 4.321928$		
$\hat{T}_3 = \hat{H}_3 - \hat{H}_4$ = 0.9791728 - 0.744458 = 0.0547148			$\hat{T}_2 = \hat{H}_2 - \hat{H}_3$ = 0.144852- 2.769779 = 0.3301422			$\hat{T}_1 = \hat{H}_1 - \hat{H}_2$ = 3.352782 - 1.744764 = 1.608018			$\hat{T}_0 = \hat{H}_0 - \hat{H}_1$ = 4.321928-3.352782 = 0.931062		
$X^2 = 1.3863KT_1$ $X^2 = 1.3863(58718)(2.100867)$ $X^2 = 19105.98018$			$X^2 = 1.3863KT_2$ $X^2 = 1.3863(58718)(0.3301422)$ $X^2 = 26873.80916$			$X^2 = 1.3863KT_1$ $X^2 = 1.3863(58718)(2.100867)$ $X^2 = 171012.1809$					
$df = c^{n-1}(c-1)^2$ $df = 20^{3-1}(20-1)^2$ $df = 361(19)^2$ = 130321			$df = c^{n-1}(c-1)^2$ $df = 20^{2-1}(20-1)^2$ $df = 20(19)^2$ = 7220			$df = c^{n-1}(c-1)^2$ $df = 20^{1-1}(20-1)^2$ $df = 1(19)^2$ =361					
p(X^2 =12836.59586, df=130321)=.99			p(X^2 = 14991.52987, df=7220)=0			p(X^2 =91851.96093, df=361)=0					

Table 34: Approval phase—Order of the process

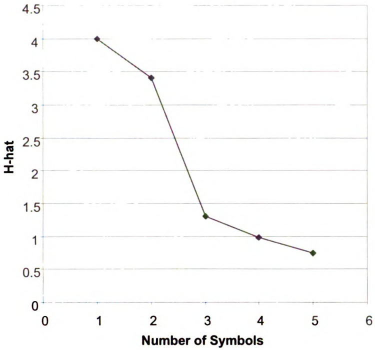


Figure 15: Scree plot for $\widehat{H}_i$ ($\hat{H}_0 = \log c, \widehat{H}_i = \hat{H}_{symbol}$) for entry phase

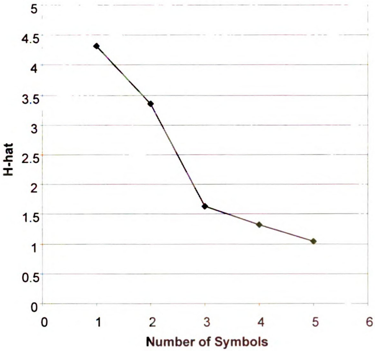


Figure 16: Scree plot for $\widehat{H}_i$ ($\hat{H}_0 = \log c, \widehat{H}_i = \hat{H}_{symbol}$) for approval phase

Phase	Group	N	Avg Auto Actions	Min	Max	Avg Length	Min	Max	Avg Auto %	Min	Max
Entry	1	194	1.69	0	2	10.75	9	14	15.53%	0.00%	18.18%
	2	562	2.43	2	5	10.13	9	20	23.89%	20.00%	38.46%
	3	591	4.62	4	6	10.14	10	14	45.50%	40.00%	54.55%
	4	653	7.29	6	10	10.07	9	13	72.38%	60.00%	100.00%
Approval	1	1030	0.58	0	3	11.62	2	45	4.48%	0.00%	9.52%
	2	526	1.96	1	4	14.88	7	31	13.15%	10.00%	15.00%
	3	595	2.15	1	4	13.30	6	23	16.16%	15.38%	18.18%
	4	701	4.74	1	13	15.21	4	37	31.95%	18.75%	78.57%

Table 35: Group statistics for visual stratification based on automation

Entry Phase p(LR = 1140.463, df =255) = 0

DF = $(T-1)(s)^r(s-1)$ (according to G+R)
T = 2 (segments)
s = 16 (number of codes)
r = 1 (order of sequence)
DF = 240
DF = 255 (according to loglin in R)
The data are judged non-stationary for the entry phase.

Approval Phase p(LR = 5247.366, df =399) = 0

DF = $(T-1)(s)^r(s-1)$ (according to G+R)
T = 2 (segments)
s = 20 (number of codes)
r = 1 (order of sequence)
DF = 380
DF = 399 (according to loglin in R)

Table 36: Omnibus test of stationarity results

Phase	Test	LR	df	p-value
Entry	Overall	3227.09	765	0.
	1 v 2	214.66	255	0.969
	1 v 3	1201.03	255	0.
	1 v 4	1921.60	255	0.
	2 v 3	1060.50	255	0.
	2 v 4	1770.55	255	0.
	3 v 4	228.23	255	0.885
Approval	Overall	6990.04	1197	0.
	1 v 2	276.46	399	1.
	1 v 3	2568.14	399	0.
	1 v 4	3946.30	399	0.
	2 v 3	2617.73	399	0.
	2 v 4	4036.11	399	0.
	3 v 4	624.95	399	0.

Table 37: Subsequent tests of homogeneity

Appendix 3: String Distance Analysis

$$A = \sum_{i=1}^{N} x_{ij} \quad B = \sum_{i=i}^{N} x_{ij}$$

$$J = \sum_{i=1}^{N} \min\left(x_{ij}, x_{ik}\right)$$

$$d_{jk} = \frac{A + B - 2j}{A + B - J}$$

Equation 5: Jaccard distance calculation

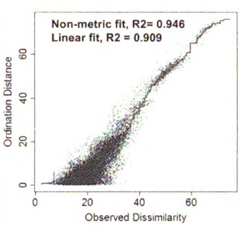
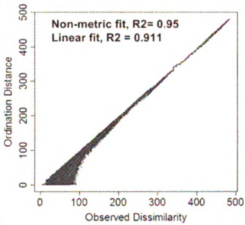
Dim	Entry	Approval
1		

Figure 17: Entry and approval Sheppard Plots, 1 through 5 dimensions, continued next two pages

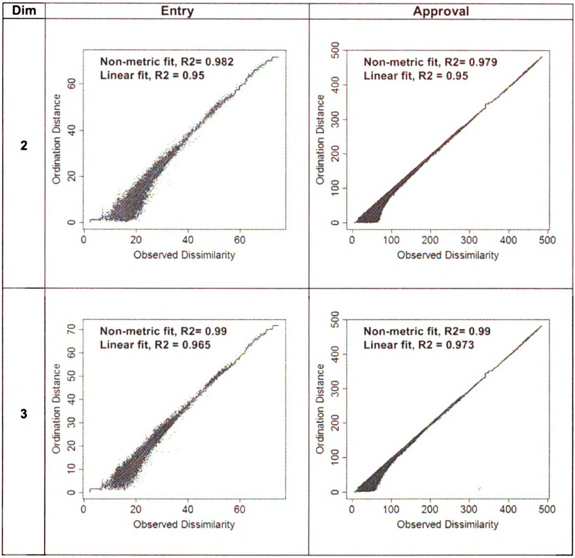


Figure 17 continued

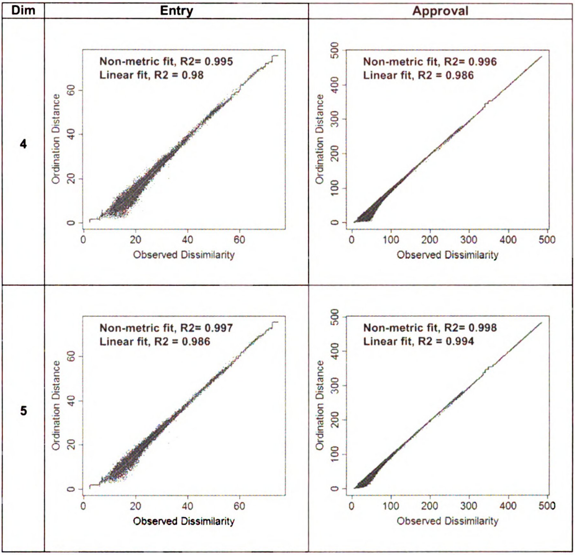


Figure 17 continued

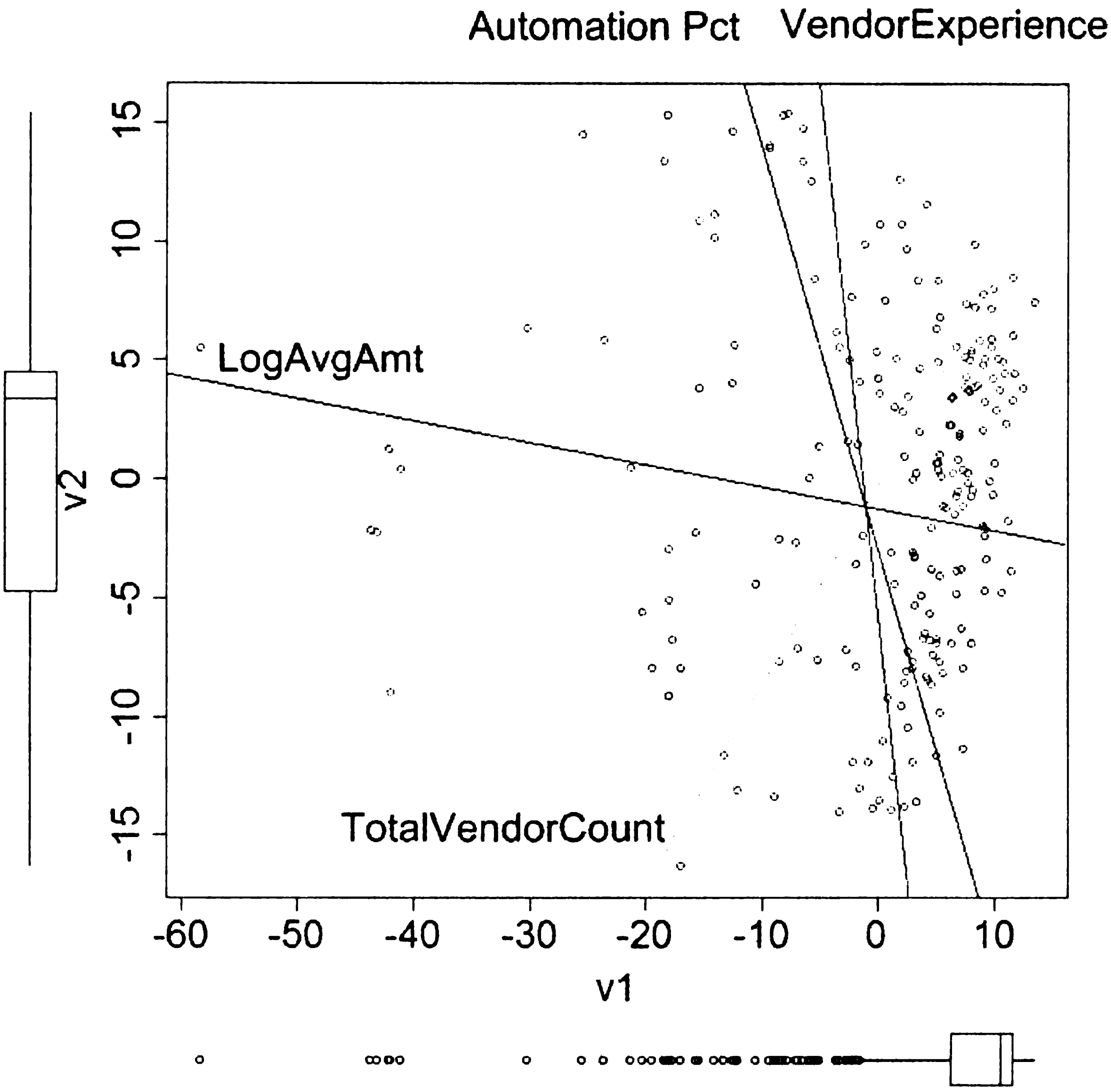


Figure 18: V1 and V2 of 3-d projection, entry phase showing lines representing the regression coefficients of variables of interest

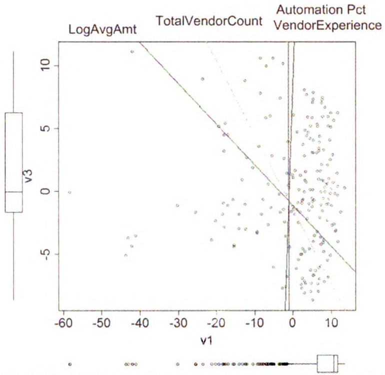


Figure 19: V1 and V3 of 3-d projection, entry phase showing lines representing the regression coefficients of variables of interest

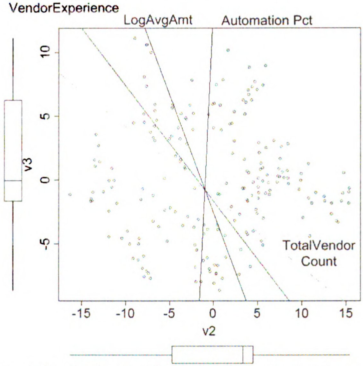


Figure 20: V2 and V3 of 3-d projection, entry phase showing lines representing the regression coefficients of variables of interest

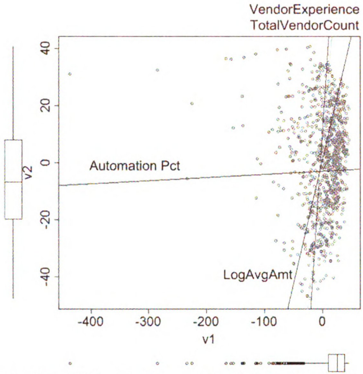


Figure 21: V1 and V2 of 3-d projection, approval phase showing lines representing the regression coefficients of variables of interest

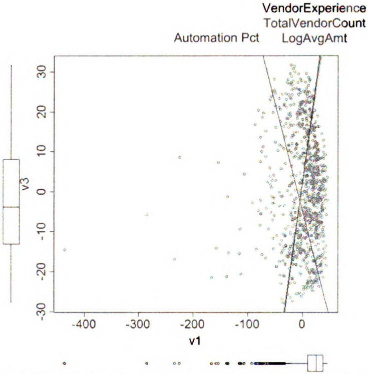


Figure 22: V2 and V3 of 3-d projection, approval phase showing lines representing the regression coefficients of variables of interest

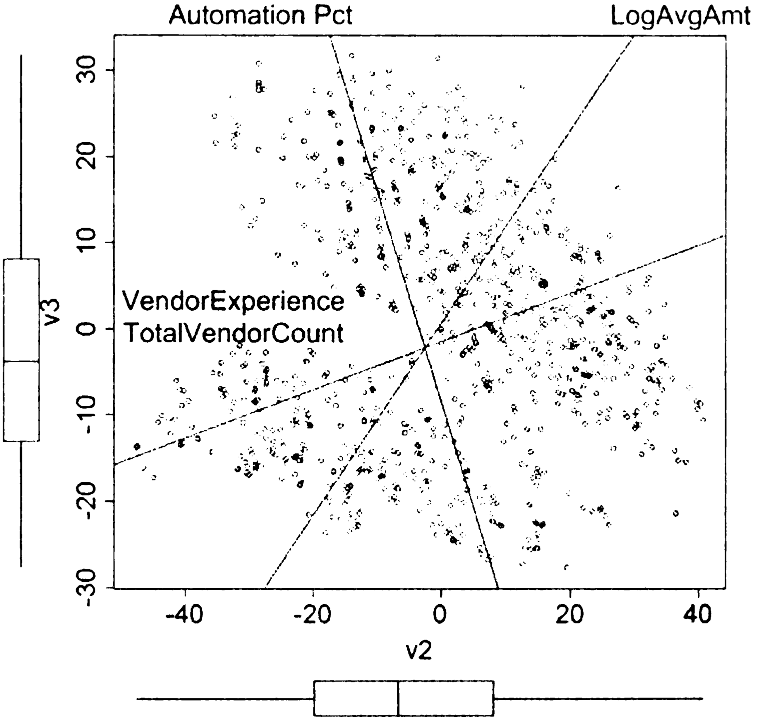


Figure 23: V1 and V2 of 3-d projection, approval phase showing lines representing the regression coefficients of variables of interest

Covariance	Automation PCT	LogAvg Amount	SC_to_AA	TotalVendorCount
AutomationPCT	0.054	16.052	45.418	87.547
LogAvgAmount	16.052	498009.694	467128.465	469830.676
SC_to_AA	45.418	467128.465	6217371.840	460210.856
TotalVendorCount	87.547	469830.676	460210.856	927401.005
v1	-0.294	2214.461	3025.708	1717.014
v2	0.547	91.851	-40.992	1477.257
v3	-0.656	195.146	-307.234	-503.161
VendorExperience	59.881	17399.337	22468.069	377922.924
	v1	**v2**	**v3**	**VendorExperience**
AutomationPCT	-0.294	0.547	-0.656	59.881
LogAvgAmount	2214.461	91.851	195.146	17399.337
SC_to_AA	3025.708	-40.992	-307.234	22468.069
TotalVendorCount	1717.014	1477.257	-503.161	377922.924
v1	46.976	3.636	5.046	-167.441
v2	3.636	46.097	0.485	1099.693
v3	5.046	0.485	18.475	-627.847
VendorExperience	-167.441	1099.693	-627.847	305797.254

Table 38: Entry covariance matrix for input, process, outcome, and automation variables

Correlation	AutomationPCT	LogAvgAmount	SC_to_AA	TotalVendorCount
AutomationPCT	1.000	0.098	0.078	0.391
LogAvgAmount	0.098	1.000	0.265	0.691
SC_to_AA	0.078	0.265	1.000	0.192
TotalVendorCount	0.391	0.691	0.192	1.000
v1	-0.185	0.458	0.177	0.260
v2	0.346	0.019	-0.002	0.226
v3	-0.657	0.064	-0.029	-0.122
VendorExperience	0.466	0.045	0.016	0.710
	v1	**v2**	**v3**	**VendorExperience**
AutomationPCT	-0.185	0.346	-0.657	0.466
LogAvgAmount	0.458	0.019	0.064	0.045
SC_to_AA	0.177	-0.002	-0.029	0.016
TotalVendorCount	0.260	0.226	-0.122	0.710
v1	1.000	0.078	0.171	-0.044
v2	0.078	1.000	0.017	0.293
v3	0.171	0.017	1.000	-0.264
VendorExperience	-0.044	0.293	-0.264	1.000

Table 39: Entry correlation matrix for input, process, outcome, and automation variables

Covariance	Automation PCT	LogAvg Amount	SC_to_AA	TotalVendorCount	v1	v2	v3	VendorExperience
AutomationPCT	0.022	-0.031	81.303	51.141	-0.807	-0.068	0.201	38.412
LogAvgAmount	-0.031	0.586	-483.426	-106.754	0.305	0.976	-1.018	-78.109
SC_to_AA	81.303	-483.426	10137830.000	184550.039	16120.018	9235.476	7180.519	139958.226
TotalVendorCount	51.141	-106.754	184550.000	557791.952	-3054.643	-3681.087	-1269.223	467752.043
v1	-0.807	0.305	16120.020	-3054.643	771.695	59.252	-3.384	-2170.481
v2	-0.068	0.976	9235.476	-3681.087	59.252	313.865	41.829	-3204.050
v3	0.201	-1.018	7180.519	-1269.223	-3.384	41.829	163.151	-1020.560
VendorExperience	38.412	-78.109	139958.200	467752.043	-2170.481	-3204.050	-1020.560	401319.356

Table 40: Approval covariance matrix for input, process, outcome, and automation variables

Correlation	AutomationPCT	LogAvgAmount	SC_to_AA	TotalVendorCount	v1	v2	v3	VendorExperience
AutomationPCT	1.000	-0.276	0.174	0.466	-0.198	-0.026	0.107	0.413
LogAvgAmount	-0.276	1.000	-0.198	-0.187	0.014	0.072	-0.104	-0.161
SC_to_AA	0.174	-0.198	1.000	0.078	0.182	0.164	0.177	0.069
TotalVendorCount	0.466	-0.187	0.078	1.000	-0.147	-0.278	-0.133	0.989
v1	-0.198	0.014	0.182	-0.147	1.000	0.120	-0.010	-0.123
v2	-0.026	0.072	0.164	-0.278	0.120	1.000	0.185	-0.285
v3	0.107	-0.104	0.177	-0.133	-0.010	0.185	1.000	-0.126
VendorExperience	0.413	-0.161	0.069	0.989	-0.123	-0.285	-0.126	1.000

Table 41: Approval correlation matrix for input, process, outcome, and automation variables

References

Abbott, A. (1983). Sequences of social events: Concepts and methods for the analysis of order in social processes. *Historical Methods, 16*(4), 129.

Abbott, A. (1990a). Conceptions of time and events in social science methods: Causal and narrative approaches. *Historical Methods, 23*(4), 140.

Abbott, A. (1990b). A primer on sequence methods. *Organization Science, 1*(4), 375-392.

Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology, 21*, 93-113.

Abbott, A., & Hrycak, A. (1990). Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. *The American Journal of Sociology, 96*(1), 144-185.

Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods Research, 29*(1), 3-33.

Agrawal, R., Gunopulos, D., & Leymann, F. (1998). *Mining process models from workflow logs*: Springer.

Agrawal, R., & Srikant, R. (1995, 1995). *Mining sequential patterns.* Paper presented at the Eleventh International Conference on Data Engineering

Alles, M. G., Brennan, G., Kogan, A., & Vasarhelyi, M. A. (2006). Continuous monitoring of business process controls: A pilot implementation of a continuous auditing system at siemens (Vol. 7, pp. 137-161): Elsevier.

Anderson, T. W., & Goodman, L. A. (1957). Statistical inference about markov chains. *The Annals of Mathematical Statistics, 28*(1), 89-110.

Ashby, W. R. (1956). Self-regulation and requisite variety. *Systems Thinking, Penguin Books, Harmondsworth.*

Ashby, W. R. (1958). Requisite variety and its implications for the control of complex systems. *Cybernetica, 1*(2), 83-99.

Ashby, W. R. (1968). Variety, constraint, and the law of requisite variety. *Modern Systems Research for the Behavioural Scientist*, 129-136.

Ashby, W. R. (1976). *An introduction to cybernetics*: Harper & Row.

Baird, D., & Weisberg, R. (1982). Rules, standards, and the battle of the forms (Vol. 68, pp. 1217–1262).

Barley, S. R. (1986). Technology as an occasion for structuring: Evidence from the observation of ct scanners and the social order of radiology departments. *Administrative Science Quarterly, 31*, 78-108.

Barley, S. R. (1990). Images of imaging: Notes on doing longitudinal fieldwork. *Organization Science, 1*(3), 220-247.

Basu, A., & Kumar, A. (2002). Research commentary: Workflow management issues in e-business. *Information Systems Research, 13*(1), 1-14.

Becker, M. C. (2004). Organizational routines: A review of the literature. *Industrial and Corporate Change, 13*(4), 643-678.

Benders, J., Batenburg, R., & van der Blonk, H. (2006). Sticking to standards; technical and other isomorphic pressures in deploying erp-systems. *Information and Management, 43*(2), 194-203.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). Discrete multivariate analysis: Theory and practice.

Board, I. S. (2002). Continuous auditing: Is it fantasy or reality? *Information Systems Control Journal, 5.*

Bollen, K. A. (1990). *Structural equations with latent variables*. New York: Wiley.

Brown, S. L., & Eisenhardt, K. M. (1997). The art of continuous change: Linking complexity theory and time-paced evolution in relentlessly shifting organizations. *Administrative Science Quarterly, 42*, 1-34.

Brynjolfsson, E., & Hitt, L. M. (1995). Information technology as a factor of production: The role of differences among firms. *Management Science, 3*(3), 183-200.

Brynjolfsson, E., & Hitt, L. M. (2000). Beyond computation: Information technology, organizational transformation and business performance. *The Journal of Economic Perspectives, 14*(4), 23-48.

Buchner, A. G., Baumgarten, M., Anand, S. S., Mulvenna, M. D., & Hughes, J. G. (1999). *Navigation pattern discovery from internet data*. Paper presented at the WEBKDD'99. from http://www.infj.ulst.ac.uk/~cbgv24/PDF/WEBKDD99.pdf.

Buckley, W. F. (1967). *Sociology and modern systems theory*. Upper Saddle River, NJ: Prentice Hall.

Carlsen, S. (1997). Conceptual modeling and composition of flexible workflow models. *Norwegian University of Science and Technology.*

Chen, M., Chen, A. N. K., & Shao, B. B. M. (2003). The implications and impacts of web services to electronic commerce research and practices. *Journal of Electronic Commerce Research 4*(4), 128-139.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: L. Erlbaum Associates

Cohen, M. D. (2007). Reading dewey: Reflections on the study of routine (Vol. 28, pp. 773): EGOS.

Cohen, M. D., & Bacdayan, P. (1994). Organizational routines are stored as procedural memory: Evidence from a laboratory study. *Organization Science, 5*(4), 554-568.

Cohen, M. D., Burkhart, R., Dosi, G., Egidi, M., Marengo, L. W., M., & Winter, S. (1996). Routines and other recurring action patterns of organizations: Contemporary research issues. *Industrial and Corporate Change, 5*, 653-698.

Compello Software, A. (2007). Retrieved 4/27/07, from http://www.compello.com/

Cook, J. E., & Wolf, A. L. (1998). Discovering models of software processes from event-based data. *ACM Transactions on Software Engineering and Methodology, 7*(3), 215-249.

Cooley, R. (2000). *Web usage mining: Discovery and application of interesting patterns from web data*. UNIVERSITY OF MINNESOTA.

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns (Vol. 1, pp. 5-32).

Cooper, A. C., & Smith, C. G. (1992). How established firms respond to threatening technologies. *Academy of Management Executive, 6*(2), 55-70.

Crowston, K. (1997). A coordination theory approach to organizational process design. *Organization Science, 8*(2), 157-175.

Culnan, M. J. (1992). Processing unstructured organizational transactions: Mail handling in the us senate. *Organization Science, 3*(1), 117-137.

Cyert, R. M., & March, J. G. (1963). *A behavioral theory of the firm*. Englewood Cliffs, NJ: Prentice-Hall.

Davenport, T. H., & Short, J. E. (1990). The new industrial engineering: Information technology and business process redesign.

David, H. A., & Nagaraja, H. N. (2004). *Order statistics*: Wiley-Interscience.

Davis, S. (1989). From future perfect: Mass customization. *Planning Review, 2*, 22.

Devaraj, S., & Kohli, R. (2003). Performance impacts of information technology: Is actual usage the missing link? *Management Science, 49*(3), 273-289.

Dewan, S., & Min, C.-k. (1997). The substitution of information technology for other factors of production: A firm level analysis. *Management Science, 43*(12), 1660-1675.

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators. *Journal of Marketing Research, XXXVII*, 269-277.

Dijkstra, W. I. L. (2001). How to measure the agreement between sequences: A comment. *Sociological Methods Research, 29*(4), 532-535.

Dijkstra, W. I. L., & Taris, T. (1995). Measuring the agreement between sequences. *Sociological Methods Research, 24*(2), 214-231.

Dunn, C. L., Cherrington, J. O., & Hollander, A. S. (2005). *Enterprise information systems: A pattern-based approach*: McGraw-Hill/Irwin.

El-Ramly, M., Stroulia, E., & Sorenson, P. (2002). *From run-time behavior to usage scenarios: An interaction-pattern mining approach* Paper presented at the Eighth ACM SIGKDD international conference on Knowledge discovery and data mining Edmonton, Alberta, Canada.

Feldman, M. S. (2000). Organizational routines as a source of continuous change. *Organization Science, 11*(6), 611-629.

Feldman, M. S., & Pentland, B. (2003). Reconceptualizing organizational routines as a source of flexibility and change. *Administrative Science Quarterly, 48*(1), 94-118.

Georgakopoulos, D., Hornick, M., & Sheth, A. (1995). An overview of workflow management: From process modeling to workflow automation infrastructure. *Distributed and Parallel Databases, 3*(2), 119-153.

Gosain, S. (2004). Enterprise information systems as objects and carriers of institutional forces: The new iron cage. *Journal of the Association for Information Systems, 5*(4), 151-182.

Gottman, J. M., & Roy, A. K. (1990). *Sequential analysis: A guide for behavioral researchers*. Cambridge: Cambridge University Press.

Grant, R. M. (1996). Toward a knowledge-based theory of the firm. *Strategic Management Journal, 17*, 109-122.

Hammer, M. (1990). Reengineering work: Don't automate, obliterate. *Harvard Business Review, 68*(4), 104-112.

Howard-Grenville, J. A. (2005). The persistence of flexible organizational routines: The role of agency and organizational context. *Organization Science, 16*(6), 618.

Jarvis, C. B., Mackenzie, S. B., & Podsakoff, P. M. (2003). A critical reivew of construct indicators and measurement model misspecifcaiton in marketing and consumer research. *Journal of Consumer Research, 30*.

Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley.

Khandwalla, P. N. (1974). Mass output orientation of operations technology and organizational structure. *Administrative Science Quarterly, 19*(74), 74-97.

Klatell, J. M. (2006, Sept. 23, 2006 =). Is mail safer since anthrax attacks? Questions remain about post office security 5 years after 5 died. Retrieved 5/30/2009, 2009, from http://www.cbsnews.com/stories/2006/09/23/eveningnews/main2036244.shtml

Koberg, C. (1988). Dissimilar structural and control profiles of educational and technical organizations. *Journal of Management Studies, 25*(2), 121.

Koestler, A. (1967). *The ghost in the machine*. London: Hutchinson.

Kohli, R., & Hoadley, E. (2006). Towards developing a framework for measuring organizational impact of it-enabled bpr: Case studies of three firms. *ACM SIGMIS Database, 37*(1), 40-58.

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey *SIGKDD Explor. Newsl, 2*(1), 1-15.

Kruskal, J. B., & Carrol, J. D. (1969). Geometric models and badness-of-fit functions. In P. R. Krishnaiah (Ed.), *Multivariate analysis* (Vol. 2, pp. 639-670). New York: Academic Press.

Kruskal, J. B., & Wish, M. (1978). Multidimensional scaling.

Langley, A. (1999). Strategies for theorizing from process data. *The Academy of Management Review, 24*(4), 691-710.

Lee, R. G., & Dale, B. G. (1998). Business process management: A review and evaluation. *Journal, Vol, 4*(3), 214-225.

Leidner, R. (1993). *Fast food, fast talk: Service work and the routinization of everyday life*: University of California Press.

Lynn, M. L. (2005). Organizational buffering: Managing boundaries and cores. *Organization Studies, 26*(1), 37.

Malone, T. W., Crowston, K., Lee, J., Pentland, B., Dellarocas, C., Wyner, G., et al. (1999). Tools for inventing organizations: Toward a handbook of organizational processes. *Management Science, 45*(3), 425-443.

March, J. G. (1991). Exploration and exploitation in organizational learning. *2*(1), 71-87.

March, J. G., & Simon, H. A. (1958). *Organizations*. New York: John Wiley and Sons.

Markus, M. L., & Robey, D. (1988). Information technology and organizational change: Causal structure in theory and research. *Management Science, 34*(5), 583-598.

Melão, N., & Pidd, M. (2000). A conceptual framework for understanding business processes and business process modelling. *Information Systems Journal, 10*(2), 105-129.

Melão, N., & Pidd, M. (2008). Business processes: Four perspectives. In M. L. Markus & V. Grover (Eds.), *Business process transformation (advances in management information systems)*: Publisher: M.E. Sharpe (April 2008).

Merton, R. (1936). The unintended consequences of purposive social action. *American Sociological Review, 1*, 894-904.

Meznar, M. B., & Nigh, D. (1995). Buffer or bridge? Environmental and organizational determinants of public affairs activities in american firms. *The Academy of Management Journal, 38*(4), 975-996.

Miller, K. D., Zhao, M., & Calantone, R. J. (2006). Adding interpersonal learning and tacit knowledge to march's exploration-exploitation model. *Academy of Management Journal, 49*(4), 709-722.

Monge, P. R. (1990). Theoretical and analytical issues in studying organizational processes. *Organization Science, 1*(4), 406-430.

Mooney, J. G., Gurbaxani, V., & Kraemer, K. L. (1996). A process oriented framework for assessing the business value of information technology. *ACM SIGMIS Database, 27*(2), 68-81.

Mukhopadhyay, T., Rajiv, S., & Srinivasan, K. (1997). Information technology impact on process output and quality. *Management Science, 43*(12), 1645-1659.

Narendra, N. C. (2004). Flexible support and management of adaptive workflow processes. *Information Systems Frontiers, 6*(3), 247-262.

Nelson, S. G., & Winter, R. R. (1982). *An evolutionary theory of economic change.* Cambridge, MA: Harvard University Press.

O'Neill, P., & Sohal, A. S. (1999). Business process reengineering: A review of recent literature. *Technovation, 19*(9), 571-581.

Oakland, J. S. (1999). *Statistical process control*: Butterworth-Heinemann Boston.

Oksanen, J. (2009a). Multivariate analysis of ecological communities in r: Vegan tutorial. Retrieved 5/11/2009, 2009, from http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf

Oksanen, J. (2009b). Package 'vegan': Reference manual. Retrieved 5/11/2009, 2009, from http://cran.r-project.org/web/packages/vegan/vegan.pdf

Orlikowski, W. J. (1992). The duality of technology: Rethinking the concept of technology in organizations. *Organization Science, 3*(3), 398-427.

Orlikowski, W. J. (1995). *Improvising organizational transformation over time: A situated change perspective*: Sloan School of Management, Massachusetts Institute of Technology.

Overby, E. (2008). Process virtualization theory and the imapct of information technology. *Organization Science, Articles in Advance*, 1-14.

Pentland, B. (1992). Organizing moves in software support hot lines. *Administrative Science Quarterly, 37*(4), 527-548.

Pentland, B. (1999). Building process theory with narrative: From description to explanation. *The Academy of Management Review, 24*(4), 711-724.

Pentland, B. (2003a). Conceptualizing and measuring variety in organizational work processes. *Management Science, 49*(7), 857-870.

Pentland, B. (2003b). Sequential variety in work processes. *Organization Science, 14*(5), 528-540.

Pentland, B., Haerem, T., & Hillison, D. (2007). *Using workflow data to explore the structure of an organizational routine.* Paper presented at the 3rd International Conference on Organizational Routines: Empirical Research and Conceptual Foundations.

Pentland, B., Haerem, T., & Hillison, D. (2009a). Comparing organizational routines as recurrent patterns of action. Unpublished Working Paper.

Pentland, B., Haerem, T., & Hillison, D. (2009b). Longitudinal endogenous changes in the performance of organizataional routines. Unpublished Working Paper.

Pentland, B., Haerem, T., & Hillison, D. (2009c). The (n)ever changing world: Stability and change in organizational routines. Unpublished Working Paper.

Pentland, B., Haerem, T., & Hillison, D. (2009d). Using workflow data to explore the structure of an organizational routine. In M. Becker & N. Lazaric (Eds.), *Organizational routines: Advancing empirical research* (pp. 47-67). Cheltenham: Edward Elgar.

Pentland, B., & Rueter, H. H. (1994). Organizational routines as grammars of action. *Administrative Science Quarterly, 39*(3), 484-510.

Peters, L., & Saidin, H. (2000). It and the mass customization of services: The challenge of implementation. *International Journal of Information Management, 20*(2), 103-119.

Petter, S., Straub, D., & Rai, A. (2007). Specifying formative constructs in information systems research. *MIS Quarterly, 31*(4), 623-656.

Poole, M. S., & Desanctis, G. (1990). Understanding the use of group decision support systems. In C. Steinfield & M. L. Markus (Eds.), *Organizations and communication technology*. Newbury Park, CA: Sage.

Rényi, A. (1953). On the theory of order statistics. *Acta Mathematica Hungarica, 4*(3), 191-231.

Sabherwal, R., & Robey, D. (1993). An empirical taxonomy of implementation processes based on sequences of events in information system development. *Organization Science, 4*(4), 548-576.

Sabherwal, R., & Robey, D. (1995). Reconciling variance and process strategies for studying information systems development. *Information Systems Research, 6*(4), 303-327.

Sankoff, D., & Kruskal, J. B. (1983). *Time warps, strings edits, and macromolecules: The theory and practice of sequence comparison*. Reading, MA: Addison-Wesley.

Schulz, M. (2008). Staying on track: A voyage to the internal mechanisms of routine reproduction. In M. Becker (Ed.), *Handbook of organizational routines*. Cheltenham: Edward Elgar.

Scott, W. R., & Davis, G. F. (2007). *Organizations and organizing: Rational, natural, and open system perspectives*. Upper Saddle River, NJ: Pearson Prentice Hall.

Shannon, C. E., & Weaver, W. (1949). The mathematical theory of communication: University of illinois press. *Urbana, 117*.

Simon, H. A. (1973). Applying information technology to organization design. *Public Administration Review, 33*(3), 268-278.

Simon, H. A. (1996). *The sciences of the artificial*. Cambridge, Massachusetts: MIT Press.

Singh, H., Pentland, B., Yakura, E., & Hillison, D. (2009). Business process management: A review and new directions.

Sorenson, O. (2003). Interdependence and adaptability: Organizational learning and the long-term effect of integration. *Management Science, 49*(4), 446-463.

Spender, J. C., & Kessler, E. H. (1995). Managing the uncertainties of innovation: Extending thompson (1967). *Human Relations, 48*(1), 35.

Thompson, J. D. (1967). *Organizations in action*: McGraw-Hill New York.

van de Ven, A. H., Angle, H. L., & Poole, M. S. (Eds.). (1989). *Research on the management of innovation: The minnesota studies*. New York: Ballinger/Harper and Row.

van de Ven, A. H., & Poole, M. S. (1990). Methods for studying innovation development in the Minnesota innovation research program. *Organization Science, 1*(3), 313-335.

van der Aalst, W. (2003). Business process management: Past, present and future.

van der Aalst, W., Desel, J., & Oberweis, A. (2000). *Business process management, models, techniques, and empirical studies*: Springer-Verlag London, UK.

van der Aalst, W., ter Hofstede, A. H. M., & Dumas, M. (2005). Patterns of process modeling. In M. Dumas, W. van der Aalst & A. H. M. ter Hofstede (Eds.), *Process-aware information systems: Bridging people and software through process technology* (pp. 179-203): Wiley & Sons.

van der Aalst, W., & van Dongen, B. F. (2002). Discovering workflow performance models from timed logs. *International Conference on Engineering and Deployment of Cooperative Information Systems (EDCIS 2002), 2480*, 45–63.

van der Aalst, W., van Dongen, B. F., Herbst, J., Maruster, L., Schimm, G., & Weijters, A. J. M. M. (2003). Workflow mining: A survey of issues and approaches. *Data & Knowledge Engineering, 47*(2), 237-267.

van der Aalst, W., & Weijters, A. J. M. M. (2004). Process mining: A research agenda. *Computers in Industry, 53*(3), 231-244.

van der Aalst, W., Weijters, A. J. M. M., & Maruster, L. (2004). Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering, 16*(9), 1128-1142.

van Driel, K., & Oosterveld, P. (2001). Nonoptimal alignment: A comment on "Measuring the agreement between sequences" By dijkstra and taris. *Sociological Methods Research, 29*(4), 524-531.

Vasarhelyi, M. A., & Halper, F. B. (1989). The continuous audit of online systems. *Artificial Intelligence in Accounting and Auditing: Knowledge Representation, Accounting Applications and the Future*, 175.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 425-478.

Wagner, H., Beimborn, D., Franke, J., & Weitzel, T. (2006). It business alignment and it usage in operational processes: A retail banking case. *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on, 8.*

Weick, K. E. (1979). *The social psychology of organizing*: Mc-Graw-Hill Publishing Co.

Weick, K. E. (1998). Introductory essay: Improvisation as a mindset for organizational analysis. *Organization Science, 9*(5), 543-555.

Weske, M., van der Aalst, W., & Verbeek, H. M. W. (2004). Advances in business process management (Vol. 50, pp. 1-8): Elsevier.

Winter, S. (1964). Economic "Natural selection" And the theory of the firm. *Yale Economic Essays, 4*, 225-272.

Yan, A., & Louis, M. R. (1999). The migration of organizational functions to the work unit level: Buffering, spanning, and bringing up boundaries. *Human Relations, 52*(1), 25-47.