Uncovering Latent Archetypes from Digital Trace Sequences: An Analytical Method and Empirical Example

Completed Research Paper

Introduction

The increasingly widespread availability of digital trace data (Hedman et al. 2013), such as email logs, website clicks, or financial transactions, provides new opportunities for researchers to understand human behaviors at a large scale (Lazer et al. 2009). More specifically, digital traces are records of activity carried out by humans or systems and captured by some form of information technology (Howison et al. 2011). Digital trace data is particularly useful for predictive models of behavioral trends (Shmueli and Koppius 2011), i.e., how a sequence of events leads to subsequent actions by an individual actor. For instance, sequences of digital traces have been used to study online platforms (Brunswicker and Schecter 2019), collaboration on Wikipedia (Lerner and Lomi 2017), and brokerage patterns in organizations (Quintane and Carnabuci 2016). The objective of this prior work was to determine *aggregate* behavioral trends, that is, to find the patterns which best described the collective behavior of the populations studied. However, this approach typically assumes total homogeneity across actors with regards to behavior over time.

In this study we focus on an alternative methodological and conceptual approach: to identify *subsets* of the population which demonstrate similar behavioral trends. The objective of this analysis would be to identify a finite set of behavioral *archetypes*, which we define as distinct patterns of action displayed by unique subsets of a population. Essentially, we assume that there is heterogeneity in how actors behave as a function of prior events, but that there are groupings or "clusters" of actors whose behavioral patterns are statistically comparable. There are some similar examples of this research design in the extant literature. Singh et al. (2011) used Hidden Markov Models (HMMs) to demonstrate that software developers may exhibit different contribution patterns based on their learning style. Another study using HMMs found that members of an online community will provide answers to questions at different rates based on their underlying motivations (Chen et al. 2017). Finally, Arazy et al. (2016) used an unsupervised clustering method to show that Wikipedia contributors will engage in different types of editing behavior based on their emergent role.

In each of these studies, an unsupervised method was applied to identify distinct behavioral patterns under different circumstances. However, there are limitations to using these techniques to identify archetypes from digital trace data. HMMs are not designed to categorize actors; rather, they use digital traces to estimate the value of some latent covariate which affects the expression of observable behaviors. Clustering techniques such as K-Means are built to carry out this categorization, but they rely upon data aggregated into a single panel. Accordingly, the sequential and temporal information available in digital trace sequences would be lost.

As an alternative, we propose a technique based on stochastic blockmodeling of relational event data (DuBois, Butts, and Smyth 2013). Stochastic blockmodels have been used to detect communities in social networks by identifying common subsets of the network (or "blocks" of the adjacency matrix) probabilistically (Karrer and Newman 2011). A relational event blockmodel applies the same logic, but uses parameterized rates of interaction within and between groups, rather than a static network (DuBois, Butts, and Smyth 2013). The objective of this paper is to develop a generalized version of the relational event blockmodel and apply it to the identification and analysis of behavioral archetypes in digital trace data.

To demonstrate the utility of the method, we present an empirical case study of information sharing patterns in small teams. With a laboratory sample of 600 participants organized into thirty teams, we demonstrate significant heterogeneity in sharing propensities among individuals. In particular, we identify two distinct sets of criteria for decision making with qualitative differences. This study makes a contribution to the literature by introducing a novel methodology for analyzing sequences of digital traces. Our approach leverages the granular information available in this data to uncover distinct patterns of behavior, thus allowing for more nuanced understandings of population dynamics. Through an empirical example, we

demonstrate the utility of the method for analyzing organizational problems. We also contribute more generally to the literature on data-driven theory development (Berente et al. 2018). The unsupervised nature of our method facilitates inductive theorizing, whereby meaningful patterns and relationships are learned through an iterative process, rather than a prior deduction.

A Model for Uncovering Archetypes

In order to identify unique behavioral archetypes from sequence data, we need a model with three key elements: (1) be able to capture the effect of prior patterns on the likelihood of a subsequent event; (2) be able to account for variability in the mechanisms over time; and (3) be able to identify heterogeneity in behavioral patterns. An appropriate framework for this problem is the relational event model (REM) (Butts 2008; Butts and Marcum 2017; Schecter et al. 2018). Relational events are single instances of an action involving a sender, receiver, and timestamp. For instance, a relational event may be a single line of text in a conversation or an edit to a software repository (Brunswicker and Schecter 2019). A relational event model is built to determine the likelihood of a relational event, or realization of a network link, based on the sequence of events that have previously transpired. Relational event models combine the analytical techniques of event history modeling (Blossfeld 2001) with the graphical or link-based representation used in social network analysis. In this way, relational event models are an ideal choice for modeling the effect of generative mechanisms on behavior (Schecter et al. 2018). Further, the path dependency of the relational event model accounts for the continuously updating values of the mechanisms derived from event sequences. To account for behavioral heterogeneity, we build on prior extensions to the core REM that uncover latent classes of actions (DuBois, Butts, and Smyth 2013; DuBois, Butts, McFarland, et al. 2013; DuBois and Smyth 2010). We use these classes of actions to delineate distinct behavioral archetypes.

Model Construction

For sake of notation, we assume that an event is a unit of information $e = (i, j, t)$ comprised of the sender i, receiver j, and time t. The set $\mathcal{A}$ describes the full sequence of events, and has cardinality N. The set of decisions made by individual i is composed of all events in which i is the sender; we denote this set as $\mathcal{A}^{(i)}$. The full sequence of decisions is $\mathcal{A} = \cup_i \mathcal{A}^{(i)}$. From this information, we thus know *who* communicated with *what* and *when*. Our objective is thus to find patterns in this sequence of events which are predictive of subsequent events; in a later section we will provide some exemplar sequence statistics.

We consider events to be arrivals from an underlying nonhomogeneous Poisson process, in which each sender-receiver pair has a unique rate (DuBois, Butts, and Smyth 2013). This rate is a function of the various generative mechanisms and the weight assigned to them according to the individual's latent decision-making criteria. In particular, we use a log-linear rate to ensure positivity.

$$\log \lambda_{ij}(t) = \beta^{(i)'} X_{ij}(t) + \varepsilon_{ij}$$

In the above equation, $\beta^{(i)}$ is a vector of weights corresponding to the criteria of individual i; $X_{ij}(t)$ is a vector of statistics corresponding to each of the relevant patterns for the pair (i, j) at time t; and ε_{ij} is an unobserved error term. Because the statistics can vary at each time point, the rate is piecewise constant – every time an event occurs which causes the statistics to change, the rate is updated accordingly. The rate may be interpreted as follows: if there is a positive weight given to a generative mechanism, then the more prevalent that mechanism is, the greater the rate will be.

To account for different underlying behavioral patterns, we assume that the weight vector $\beta^{(i)}$ comes from one of a finite set of vectors, plus some unobserved heterogeneity. We model the weight vector as follows:

$$\beta^{(i)} = \sum_{p \in P} \beta_p z_{ip} + \varepsilon_i$$

Here, each z_{ip} is a binary variable, with a value of 1 indicating individual i employs archetype p. The vector β_p is the weights assigned under archetype p, and ε_i is a vector of unobserved deviations of individual i. We assume that $\sum_{p \in P} z_{ip} = 1$, i.e., each person has one distinct decision making approach. Because these indicators are unobservable, they must be estimated empirically.

Given the model ingredients we have defined, we can construct the probability of an observed event. There are two components to be modeled: the likelihood of the sender selecting the receiver, and the likelihood of the time having elapsed (Brandes et al. 2009). To determine the probability of i sending information to j, we utilize the fact we established previously that all possible events are governed by an underlying nonhomogeneous Poisson process. Events with higher rates are expected to occur more frequently and vice versa. These rates are conditionally independent of one another given the prior sequence. Accordingly, the likelihood of a given event occurring is equal to the ratio of rates (Butts 2008). Put another way, the event with the highest rate or fastest expected arrival will have the highest probability of occurring next. The choice process thus follows a multinomial distribution, with all potential recipients comprising the state space $\mathcal{R}^{(i)}$ (DuBois, Butts, and Smyth 2013; DuBois, Butts, McFarland, et al. 2013; Stadtfeld and Block 2017). Thus, the selection probability is given as:

$$p(i \to j, t) = \frac{\lambda_{ij}(t)}{\sum_{l \in \mathcal{R}^{(i)}} \lambda_{il}(t)}$$

For the time between two events, $t_m - t_{m-1}$, the interval follows an exponential distribution with mean equal to the sum of all event rates. This fact follows directly from properties of Poisson process waiting times. Therefore, the probability of observing a particular time interval is equal to:

$$p(t_m - t_{m-1}) = \left(\sum\nolimits_{j \in \mathcal{R}^{(i)}} \lambda_{ij}(t_m)\right) \exp\left(-(t_m - t_{m-1}) \sum\nolimits_{j \in \mathcal{R}^{(i)}} \lambda_{ij}(t_m)\right)$$

Combining these two elements, we may produce the likelihood of a given event in the sequence. To compute the entire likelihood function, we take the product of each event probability (Brandes et al. 2009; Butts 2008).

Model Inference

Inference must be conducted to determine the weight parameters β as well as the latent assignment variables z. To identify their values, we apply Bayes' rule to determine the conditional likelihood of each variable, given the remaining variables. We first define the posterior likelihood function for the latent variables (DuBois, Butts, and Smyth 2013):

$$\begin{aligned} p(z_{ip} \mid \mathcal{A}^{(i)}, \beta) &= \frac{p(\mathcal{A}^{(i)} \mid z_{ip}, \beta) p(z_{ip}, \beta)}{p(\mathcal{A}^{(i)}, \beta)} \\ &\propto p(\mathcal{A}^{(i)} \mid z_{ip}, \beta) \\ &= \prod_{m=1}^{M} \lambda_{ij_m}(t_m) \prod_{j \in \mathcal{R}^{(i)}} \exp\left(-(t_m - t_{m-1}) \lambda_{ij}(t_m)\right) \\ &= \prod_{m=1}^{M} \exp\left(\beta_p X_{ij_m}(t_m)\right) \prod_{j \in \mathcal{R}^{(i)}} \exp\left(-(t_m - t_{m-1}) \exp\left(\beta_p X_{ij}(t_m)\right)\right) \end{aligned}$$

The above equation can be interpreted as: the likelihood of i relying on weight vector p is proportionate to the likelihood of observing i's decisions $\mathcal{A}^{(i)}$ given their internal criteria and the corresponding weights. Note that here we assume an uninformative prior for z, though that assumption can be adjusted. The likelihood of observing the sequence of M decisions is equal to the product of each event's probability, as defined previously. Essentially, the assignment that is most likely for i is that which makes their sequence of decisions most probable.

To conduct inference on the weight parameters β, we consider all decisions made by individuals using the same criteria. Accordingly, the posterior likelihood function is defined as follows:

$$\begin{aligned} p(\beta_p \mid \mathcal{A}, z) &= \frac{p(\mathcal{A} \mid \beta_p, z) p(\beta_p, z)}{p(\mathcal{A}, z)} \\ &\propto p(\mathcal{A} \mid \beta_p, z) \end{aligned}$$

$$= \prod_{n=1}^{N} \lambda_{i_n j_n}(t_n)^{z_{i_n p}} \prod_{(i,j)\in\mathcal{R}} \exp\left(-(t_n - t_{n-1})\lambda_{ij}(t_m)\right)$$
$$= \prod_{n=1}^{N} \exp\left(\beta_p' X_{i_n j_n}(t_n)\right)^{z_{i_n p}} \prod_{(i,j)\in\mathcal{R}} \exp\left(-(t_n - t_{n-1}) \exp\left(\beta_p' X_{i\,j}(t_n)\right)\right)$$

Here, the likelihood function is equal to the probability of the sequence of decisions made by individuals using weight vector p, multiplied by the probability of the observed time intervals. Again, we assume an uninformative prior. Note that we use the assignment variable $z_{i_n p}$ as an indicator; that way, we can account for changes to the rate function which may occur due to actions by those not in assignment p (Butts and Marcum 2017; DuBois, Butts, and Smyth 2013). This expression implies that the weight vector β_p that has the highest probability will maximize the likelihood of observing the decisions made by individuals using criteria p.

To recover the parameters from our model, we apply an expectation-maximization (EM) algorithm to iteratively make class assignments and update the model parameters (Dempster et al. 1977). The steps can be summarized as follows:

1. For each individual, assign them to a group that makes their decisions most likely.
2. For each group of individuals, fit a weight vector that makes their collective decisions most likely.
3. Return to Step 1 and iterate until convergence.

By design, the EM-algorithm should converge to an optimal solution to the inference problem. With this procedure, we are able to determine the assignments of individuals to classes, as well as the selection of parameters for each strategy.

Empirical Illustration: Information Sharing in Teams

The success of face-to-face work teams (Mesmer-Magnus and DeChurch 2009) and virtual teams (Mesmer-Magnus et al. 2011) hinges on the ability of individuals to share information amongst themselves, and subsequently synthesize it in a meaningful way (Robert et al. 2008). Indeed, a high level of information sharing indicates that the requisite item was provided in a timely manner throughout the team's work cycle, "thereby enabling groups to reach higher quality solutions that could be reached by any one individual" (Mesmer-Magnus and DeChurch 2009, p. 535). When individuals share information, they broaden the information space available to the team, i.e. illuminating relevant facts, as well as enhance the potential for meaningful solutions or outcomes (Shore et al. 2015). Through frequent communication and information sharing, team members also develop an awareness of where expertise is located, who is an accurate source of knowledge, and who has access to other sources of information (Choi et al. 2010; Kanawattanachai and Yoo 2007; Leonardi 2015). This behavior over time allows for more effective sharing in the future, and subsequently higher levels of performance (Kanawattanachai and Yoo 2007).

Because information sharing is an integral component of team success, it is important to understand how information is actually transferred from one individual to another. When individuals do choose to share information and collaborate, they must determine where to route that information so that it reaches the intended recipient. However, they may not possess the meta-knowledge necessary to send information directly (Faraj and Sproull 2000; Leonardi 2015). Thus, we expect members to rely on a set of heuristics to make their decisions. In other words, instances of individuals sharing information are driven by a mix of behavioral trends and cognitive factors. Further, the role each of these effects has on a person's actions is unique to that individual. For instance, an individual may choose to share information with a frequent informal communication partner. On the other hand, a separate individual may try to share information as broadly and frequently as possible. To uncover these effects and their influence on decision-making, we shift our focus to *how* the process of transferring information occurs.

We draw on prior literature to build a picture of the information sharing process for members of teams. In particular, we focus on ad hoc virtual teams, or those without significant development of transactive memory systems (Faraj and Xiao 2006; Majchrzak et al. 2007; Majchrzak and Malhotra 2016). This distinction is important because individuals who choose to share information must rely on a set of behavioral patterns and experiences, rather than meta-knowledge of who knows who and who knows what

(Faraj and Sproull 2000). First, we argue for a dynamic perspective on information sharing in additional to a more conventional structural approach. Second, we describe the various behavioral patterns which may drive information sharing actions. We then apply our method to identify the distinct archetypes of sharing behaviors.

Information Sharing as a Temporal Process

The likelihood of an individual contributing knowledge to an organization or community can be affected by several factors, including social capital, group norms, and feelings of self-efficacy (Chiu et al. 2006; Kankanhalli et al. 2005; Wasko and Faraj 2005). As Wasko and Faraj (2000) point out, the motivations for exchanging knowledge vary based on how information is defined by the organization; when information is an object, it is exchanged through interactions with others. In this case, individuals are motivated by self-interest, e.g. gaining rewards or fulfilling obligations.

While there are various potential motivations, the actual factors underlying behavior are only expressed over time as people make decisions and engage in group processes. Information sharing occurs during the process of team coordination, or the interactions between individuals that manage resources and expertise dependencies. Coordination plays a vital role in effective information sharing by identifying who knows what, who needs what, and mobilizing the available resources (Faraj and Sproull 2000). The coordination of expertise in a team has both structural and temporal elements. The structural or configural view of coordination posits that interactions follow distinct patterns, and that there is a relationship between these structures and organizational outcomes (Kudaravalli et al. 2017; Kudaravalli and Faraj 2008). To determine how collaboration is structured, it is useful to represent the team as a network with individuals as nodes and links between the nodes signifying joint taskwork or coordination (Crawford and LePine 2013). From this perspective, an individual may choose to share information based on their local network, i.e. those they have ties with, or may direct information towards central actors. Further, individuals may delineate between informal communication networks and expertise networks, and share information with those whom they communicate with on a more general level (Sosa et al. 2015).

Coordination is temporal in that structures are not constant over time; rather, they emerge as a result of repeated actions (Kozlowski and Klein 2000; Kudaravalli et al. 2017). The notion of emergence in and of itself implies dynamism, where a series of actions, interactions, or events coalesces into a state or property of the group. Following this same logic, (Faraj and Xiao 2006) describe coordinated action as a temporal unfolding of events, and coordination mechanisms define how these events proceed. When we focus on process, rather than structure, we emphasize the temporal relationships between events and how those relationships shift (Van de Ven and Poole 2005). In the context of information sharing, the decision to transmit information from one individual to another is a distinct activity that occurs at a specific point in time, and occurs with the overarching structure providing context. The temporal perspective shifts the focus from what structures facilitate information sharing, to how individuals actually engage in sharing. Accordingly, events occur in the context of the coordination network, and can subsequently reshape that network.

Factors Influencing Information Sharing

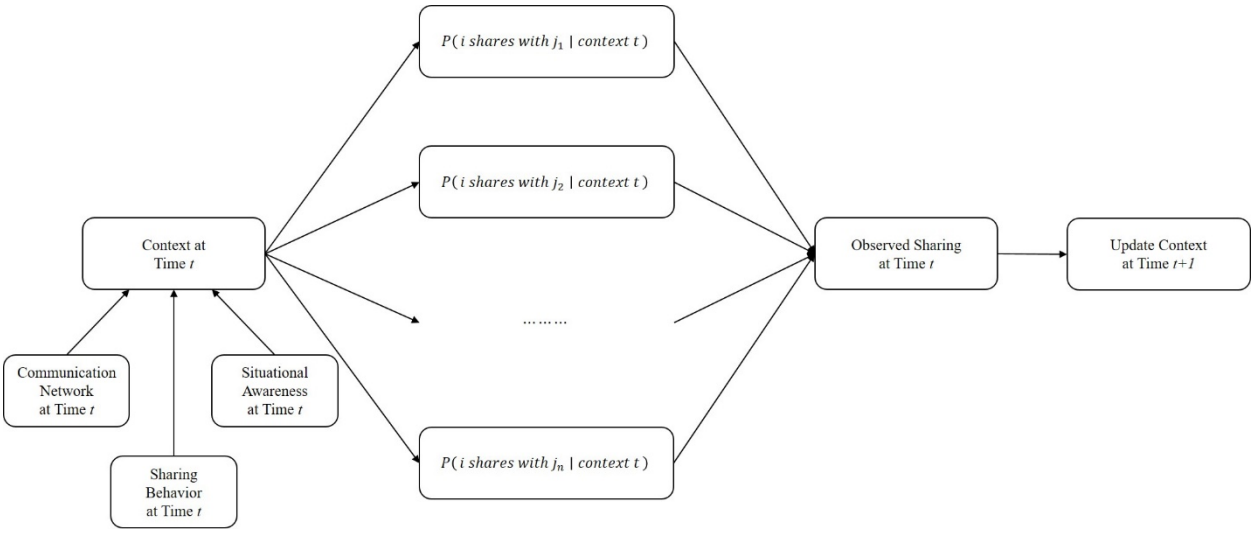
We proceed to describe three types of mechanisms that drive information sharing: tendencies derived from the communication network, tendencies derived from prior sharing behavior, and tendencies derived from individual awareness of the situation. The influence each of these factors has on the propensity to share information will be a characteristic of an individual and represent the norms governing their sharing practices (Bock et al. 2005). We summarize the architecture of an information sharing decision in Figure 1.

Communication Network

Communication network effects describe the impact of informal communication networks on the tendencies for individuals to share information (Reagans and McEvily 2003). Informal communication patterns are distinct from formal task interdependencies, and form as a result of attempts to coordinate (Sosa et al. 2015). We focus on four generative mechanisms of information sharing based on network structure: communication frequency, reciprocity, activity, and popularity. The first network mechanism is communication frequency, which refers to the tendency for individuals to send information to their more

regular contacts. Likewise, our second mechanism, communication reciprocity, refers to the act of an individual sharing information with those who contact him more frequently. Together these patterns form a dyadic approach to information sharing. Having frequent contact between members can reduce the cost associated with transferring information (Hansen et al. 2005; Reagans and McEvily 2003), and in general the degree of interconnection in the organization should promote greater sharing of knowledge (Tsai 2002). Repeated contact will make sharing information faster and more efficient, but at the cost of tie maintenance (Hansen 2002). Consequently, individuals may rely on a dyadic strategy to promote ease of transferal, but others may prefer to rely on weaker, more ephemeral ties (Granovetter 1973, 1983).

Figure 1. Architecture of sharing decision



Our next two mechanisms, communication activity and popularity, describe the tendency for individuals who are most active (many outbound messages) or most popular (many inbound messages) in the communication network to also share the most information. Essentially, one who has the greatest number of connections will have greater opportunity to transfer information due to their broader potential range (Reagans and McEvily 2003). These individual-level mechanisms could come about because of simple activity levels – i.e. one person communicates very frequently and subsequently passes a higher volume of information – or because a central individual may have a somewhat higher awareness of who needs what intelligence (Leonardi 2015).

Sharing Behavior

Sharing behavior effects describe the impact of prior sharing patterns on future transmission decisions. These interactions are distinct from other forms of communication in that they are directly related to technical details needed by the team (Kudaravalli et al. 2017; Sosa et al. 2015). These explanations include prior transference frequency, sharing reciprocity, total sharing activity, and sharing popularity. Prior transference frequency encapsulates the tendency towards inertia, where one individual tends to send information to the same set of partners. Alternatively, sharing reciprocity describes a tendency for individuals to share information with others who have previously sent them information (Faraj and Johnson 2011). Both inertia and reciprocity are indicative of a reliance on well-developed pathways and existing social capital (Chiu et al. 2006). These mechanisms mirror the patterns of frequency and reciprocity in the informal communication network.

Our third pattern, prior sharing activity, encapsulates the tendency for an individual to maintain or expand their prior rate of sharing behavior. For example, an agent may determine that an optimal strategy is to share information as frequently as possible, or to accelerate their rate of sharing over time. Finally, the fourth mechanism of sharing, popularity, is indicative of a tendency to share more frequently when more information has been received. This behavior is analogous to the pattern of indirect reciprocity found in online communities (Faraj and Johnson 2011). In other words, as an individual collects more knowledge,

their propensity to share increases. These two effects collectively describe the tendency for information to flow through individuals who are focal in the overall sharing process.

Situational Awareness

The final set of generative mechanisms we consider are based on an individual's awareness of the task parameters and the team's progress towards their goal (Marks et al. 2001). Specifically, we focus on the impact of time on decision making, though other task-specific behaviors are possible. Awareness of time constraints encapsulates an individual's reaction to the task deadline. An individual who is more aware may act proactively and share information quickly after receiving it. Alternatively, someone who is not motivated by time constraints will pass information along at random intervals, or perhaps even wait until the last minute to disseminate information. Finally, the effect of time is likely non-linear; for example, an actor may increase their rate of sharing over time, but this acceleration might taper off as the project moves into a later stage. Thus, we also consider quadratic effects of time.

Methods

Data

We collected data through a series of experiments in which participants had to complete an information sharing task in a virtual team environment. The sample is composed of 600 unique individuals organized into thirty virtual teams; all thirty teams accomplished the task according to the same conditions and parameters. Participants were recruited at a Midwestern US university and participated in this study in exchange for either research credit or $35. Individuals reported to a laboratory in groups of twenty, forming a single virtual team, and each group was conducted in a separate two-hour session. The session consisted of pre and post-game surveys, a twenty-minute practice mission, and a forty-minute performance mission. For the purpose of this study we only consider data from the performance mission component. Participants were randomly assigned to one of four functionally equivalent units within the virtual organization, and to a specific role within their component. Each participant was seated at an individual workstation, and performed the task using a laptop computer.

In each experimental session, the teams participated in a computer-based simulation game that entailed guiding a humanitarian aid convoy through dangerous territory – this task was considered the primary objective for the team. Each of the four units were responsible for clearing targets within one quadrant of the game map. The convoy could not progress through the map until the obstacles in its path were eliminated. In order to clear a target on the map, one member of the team – a *reconnaissance officer* – had to identify an obstacle by flagging the correct portion of the grid. If the recon officer chose an incorrect space on the grid, the target would remain. Once the obstacle is correctly identified, a *field specialist* on the component team could then neutralize the threat. Threats were classified as either "insurgent" or "IED" and the two threat types each had a recon officer and field specialist assigned to them. All four units also had a designated navigator who would help determine the path of the convoy.

In total, all four quadrants have thirty-two threats located inside them. However, each five-person unit is only given the locations of eight of those targets. The remaining twenty-four locations are evenly distributed among the other three groups. Each unit therefore possessed thirty-two pieces of intelligence, but twenty-four needed to be distributed to the other fifteen members of the team. The intelligence was provided randomly to participants on a sheet of paper at the beginning of the study. While each unit had to clear their own obstacles, the participants were judged on how far the convoy was moved. As such, there was a clear performance incentive to share the pieces of information that the individuals did not need. All information had to be shared through direct communication; there was no distinct game feature for sharing intelligence. In Figure 2 we provide a sample of the intelligence provided to the participants; note that participants are not given the target unit, and each participant is given unique information.

Participants wore headsets and communicated with one another through Skype. Each player had an anonymous handle that corresponded to their five-person unit. The handles indicated whether a player was a recon officer or field specialist; however, a player's threat assignment was not part of the naming scheme. Thus, players had some knowledge as to where intelligence should be routed, but that knowledge was

incomplete. Team members were allowed to choose between text-messaging through Skype chat or video calls with all other participants.

Figure 2. Sample intelligence document

Player	Source Unit	Object	Cell	X Coordinate	Y Coordinate	Target Unit
Dragon Recon	Caspia	Barrel	A2	418	444	Caspia
Dragon Recon	Caspia	Barrel	A2	121	70	Caspia
Dragon Recon	Caspia	Tire Fire	A9	149	269	Baltica
Dragon Recon	Caspia	Sedan	C1	189	210	Caspia
Dragon Recon	Caspia	Sedan	C6	352	41	Baltica
Dragon Recon	Caspia	Tire Fire	C6	345	191	Baltica
Dragon Recon	Caspia	AFV	H2	292	397	Atlantica
Dragon Recon	Caspia	RPG	H2	186	153	Atlantica
Dragon Recon	Caspia	Tent	J7	325	254	Pacifica

Notes. The target unit is not provided to the players; we include it here for illustration.

We collected a full time-stamped transcript of all communication for each session, which provides us with data on who said what to whom at what time; our final data form is that of a transcript, with each row containing sender, receiver, time, and message. Manual coders went through each line of communication and marked lines that contained pieces of intelligence. Further, the coders gave a unique identifier to each target, so the accuracy of the information sharing could be assessed. A sample of our data format and coding is given in Figure 3.

Figure 3. Sample communication excerpt coded for information content

Time	Sender	Receiver	Message	Information
20:30:30	Player 1	Player 2	I have info on a threat	N
20:30:42	Player 2	Player 1	What is it?	N
20:31:07	Player 1	Player 2	Two actually	N
20:32:26	Player 1	Player 2	AFV, Cell J2, X: 431 Y: 138	Y (ID: 48)
20:32:44	Player 1	Player 2	RPG, Cell H2, X:186, Y: 153	Y (ID: 21)

Notes: Each message was identified as containing information or not (Y/N). Coders also gave a unique ID (1-128) to each of the distinct pieces of information.

In total, we observed 118,333 messages amongst the 600 participants. Of those messages, 3,923 made direct mention of a piece of intelligence.

Measures for Sharing Patterns

To compute measures for each of our identified mechanisms, we converted our coded communication transcripts into relational event sequences. From these sequences, we computed two arrays, U and V, which were weighted adjacency matrices with value at each point in time. The entry (i, j) at time t of $U(t)$ is represented as u_{ijt}, and is equal to the number of messages i has sent j up to time t. Likewise, the entry (i, j) at time t of $V(t)$ is represented as v_{ijt}, and is equal to the number of messages i has sent j up to time t that contain coordinate information. Accordingly, $u_{ijt} \geq v_{ijt}$ for all i, j, t. Using these arrays, we can compute statistics representing the mechanisms at every point in time, for every feasible pair of people. For our measures regarding awareness of time, we simply included a metric for time elapsed in the mission. In Table 1 we list our variables and relevant formulae.

Analysis of Behaviors

To identify the unique patterns exhibited by the participants in our study, we used an unsupervised approach to determine the optimal combination of variables and groups. In particular, we fit the aforementioned model with a variety of parameter combinations – network terms only, prior sharing only, situational awareness only, pairwise combinations, and the full model – and a range of groups (i.e. P = 1, 2,

3,…). We assessed model quality using the log-likelihood, as well as the Bayesian Information Criterion (BIC) in order to avoid overfitting. The best combination of parameters and groups will have the greatest log-likelihood and smallest BIC value. If these measures indicate that a multi-group solution is optimal (i.e. P > 1), then we may conclude that there are distinct sets of decision-making criteria that govern sharing behavior.

Table 1. Key behavioral mechanisms for REM analysis

Variable Name	Description	Formula
Communication Frequency	The number of messages sent to another individual	$x_1(i,j,t) = u_{ijt}$
Communication Reciprocity	The number of messages received from another individual	$x_2(i,j,t) = u_{jit}$
Communication Activity	The number of messages an individual has sent in the past	$x_3(i,j,t) = \sum_k u_{ikt}$
Communication Popularity	The number of messages an individual has received in the past	$x_4(i,j,t) = \sum_k u_{kit}$
Sharing Frequency	The volume of information sent to another individual	$x_5(i,j,t) = v_{ijt}$
Sharing Reciprocity	The volume of information received from another individual	$x_6(i,j,t) = v_{jit}$
Sharing Activity	The volume of information an individual has sent in the past	$x_7(i,j,t) = \sum_k v_{ikt}$
Sharing Popularity	The volume of information an individual has received in the past	$x_8(i,j,t) = \sum_k v_{kjt}$
Time	The time elapsed in the mission	$x_9(i,j,t) = t$

Results

Following our analysis procedure, we tested a variety of models with different sets of parameters and a range of groups. We identify the best fitting model as one with two groups and all parameters included; the log-likelihood for P = 2 was -36,247, while the log-likelihood for P = 1 was -36,550. For P = 3, the log-likelihood marginally improved, but the BIC value did not, suggesting the additional parameters did not significantly improve the model. Thus conclude that there are two dominant archetypes in information sharing behavior. To determine what factors make up these approaches, we examine the parameter values for this best model. The results are presented in Table 2.

From Table 2 we observe that some parameters have consistent effects (i.e., sign and significance the same across archetypes) and others have varied effects. Further, the effect sizes vary significantly across archetypes for many of the mechanisms we tested. We computed the difference in parameters for each statistic, and computed the pooled standard error of the difference. A useful way to interpret the differences in effects is to compute odds ratios, which are equal to $\exp(\theta)$. In other words, for every additional event (a unit increase in the statistic), the odds ratio would give the relative odds for members of one archetype to share, relative to the others.

Table 2. Parameter estimates for two-archetype solution

	Archetype 1	**Archetype 2**	**Difference**	**Odds Ratio**
Variable	Coef (SE)	Coef (SE)	Coef (SE)	exp(Coef)
Rate	-11.0739**	-11.2098**	0.136	1.146
	(0.0739)	(0.0691)	(0.1012)	
Communication Frequency	0.9599**	1.6294**	-0.6695**	0.512**
	(0.1452)	(0.1222)	(0.1898)	
Communication Reciprocity	0.6715**	-0.6004**	1.272**	3.568**
	(0.1224)	(0.1171)	(0.1694)	
Communication Activity	7.3262**	3.0335**	4.2927**	73.164**
	(0.5313)	(0.8162)	(0.9739)	
Communication Popularity	-9.3328**	0.0505	-9.3833**	0.000**
	(0.8089)	(1.0022)	(1.2879)	
Sharing Frequency	2.7123**	2.8538**	-0.1416	0.868
	(0.0670)	(0.0557)	(0.0871)	
Sharing Reciprocity	0.1269	1.0565**	-0.9296**	0.395**
	(0.0996)	(0.0886)	(0.1333)	
Sharing Activity	1.7554**	2.8168**	-1.0613**	0.346**
	(0.1013)	(0.1058)	(0.1465)	
Sharing Popularity	3.1247**	2.108**	1.0167*	2.764*
	(0.2215)	(0.2692)	(0.3486)	
Time	3.2403**	2.2639**	0.9764*	2.655*
	(0.3189)	(0.2657)	(0.4150)	
$Time^2$	-3.8497**	-3.1353**	-0.7143	0.490
	(0.4252)	(0.3279)	(0.5369)	
N	214	254		
Deviance	32,908	39,946		

*Note: Significance code * $p < 0.01$, ** $p < 0.001$*

Based on the computed ratios, we observe that individuals following Archetype 1 are significantly less likely to share information with those whom they've *sent* more messages to (OR = 0.512), but are more likely to share information with those whom they've *received* more messages from (OR = 3.568). Further, sending more messages prior tends to increase the rate of sharing for Archetype 1, while receiving a high volume of messages tends to make it decrease.

In terms of prior sharing behavior, we find that individuals following Archetype 2 are more likely to engage in direct reciprocity (Faraj and Johnson 2011), i.e., sharing information with those who shared with them prior (OR = 0.395). Further, when those individuals share more information, they tend to increase their rate of sharing relative to Archetype 1 (OR = 0.346). However, individuals following Archetype 1 are more likely to engage in indirect reciprocity (Faraj and Johnson 2011), i.e., they share more as they receive more information generally (OR = 2.764).

Finally, we find that with Archetype 1, individuals tend to share information later (OR = 2.655), but there is no difference in the quadratic effect. In summary, the first type of individual tends to share information after having sent more messages and received more information, and tend to target those who communicated with them directly. Alternatively, the second type of individual tends to share information earlier and more frequently, and will send information to those who share with them first.

Discussion & Conclusions

In this study we introduced a method for uncovering latent behavioral archetypes within digital trace data, drawing upon prior work on stochastic blockmodeling (Karrer and Newman 2011), relational event models (DuBois, Butts, and Smyth 2013), and discrete choice models (McFadden 1974). The proliferation of digital traces prompts the need for new analytical models that leverage the granular, temporal data collected through information technologies (Howison et al. 2011). In particular, we attempt to deconstruct large sequences and identify unique sets of patterns formed by subsets of people. By understanding how past behaviors influence future actions, it is possible to uncover the latent tendencies of each actor in a population. Essentially, each individual falls into a category that describes the overall trends in their actions; though these classes cannot be observed directly, the categories can be inferred by delineating distinct action patterns (Chen et al. 2017; Singh et al. 2011). As such, we expect that every individual will have a unique behavioral "signature" that describes their pattern of activities – in our case, how, when and with whom they choose to share information. Thus, our method provides a framework for inductive analysis of human behavior, and supports theory building through iterative discovery (Berente et al. 2018).

Our empirical example also makes a contribution to the literature on team dynamics and information sharing. The event-based approach we applied is a natural extension of prior work on expertise coordination and knowledge sharing in virtual teams and online communities. Indeed, much of the extant literature emphasizes the role of interactions in shaping coordination (Faraj and Sproull 2000; Faraj and Xiao 2006; Kudaravalli et al. 2017). However, in these studies interactions tend to be compressed into static network configurations (Kudaravalli et al. 2017; c.f. Kudaravalli and Faraj 2008), thereby losing the granularity of data at the event level (see Quintane et al. 2014). Like coordination and collaboration, information sharing is a process composed of interactions between team members that unfold over time. Thus, we applied a framework that explicitly focuses on events and the temporal relationships between events in a sequence. This shift allows us to focus on when and why an action took place, and allows us to differentiate sequences, even if they start and end at the same point. For example, two individuals may share the same amount of information with the same people. However, one of these two conducts all of their sharing *before* establishing a relationship through informal communication, while the other builds strong ties first. These two patterns of behavior are distinct, however they could not be captured by a purely compositional or configural perspective (Schecter et al. 2018). Considering these two descriptions, we argue that our findings demonstrate the utility of the dynamic approach. Regardless of outcomes, we observe that individuals engage in two patterns of behavior that are not only quantitatively different, but also qualitatively different. This finding further adds to the recognition that processes – *how* things happen – are distinct from structures (Quintane and Carnabuci 2016). As such, this line of reasoning and the accompanying methodology has a significant potential for future research.

There are some limitations to the proposed methodology and the results of empirical case study, as well as potential directions for further research. Our model implicitly interprets behaviors as conscious decisions made by boundedly rational actors; of course, this implies that actors are able to accurately assess the state of the system with regards to prior events. This assumption may be difficult in large online settings with rapid updates. Further, while we do assume heterogeneity among actors, we are still collapsing individuals into discrete categories, and it is potentially impossible to determine how many categories is "correct." Indeed, the utility of delineating the population hinges largely on how distinct the groups are, and how meaningful the differences in behavior are. Accordingly, when using this method researchers should take care to qualitatively justify the discovered archetypes. Finally, our model does not explicitly account for heteroscedasticity, i.e., behaviors within a group changing over time. However, the sufficient statistics could be operationalized to account for the progression of time. A fruitful direction for future work could be the exploration of how subsets of the population change their behaviors over time.

With regard to the empirical example, the virtual teams were made up of students, and these participants had no previous experience working with one another, and had no expectation of working together in the

future. These conditions are not realistic in most work environments. However, there are also advantages: the laboratory setting allowed us to specifically track the sharing of tangible intelligence in real time and in relation to communication behavior. Future research should extend these findings to actual organizations. Additionally, while we did identify the informational content of each message, we did not account for other semantic qualities such as affect or tone. To the extent that we were able, we attempted to filter out references to information that were superfluous, such as inquiring about the state of a cell or checking to see if the obstacle has been cleared. These interactions, though they contained information, are not sharing in the context of what we are analyzing. While we believe that this analysis would be a fruitful direction for future research, it is beyond the scope of this study.

References

Arazy, O., Daxenberger, J., Lifshitz-Assaf, H., Nov, O., and Gurevych, I. 2016. "Turbulent Stability of Emergent Roles: The Dualistic Nature of Self-Organizing Knowledge Coproduction," *Information Systems Research* (27:4), pp. 792–812. (https://doi.org/10.1287/isre.2016.0647).

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

Blossfeld, H.-P. 2001. *Techniques of Event History Modeling: New Approaches to Casual Analysis*, Psychology Press.

Bock, G.-W., Zmud, R. W., Kim, Y.-G., and Lee, J.-N. 2005. "Behavioral Intention Formation in Knowledge Sharing: Examining the Roles of Extrinsic Motivators, Social-Psychological Forces, and Organizational Climate," *MIS Quarterly*, pp. 87–111.

Brandes, U., Lerner, J., and Snijders, T. A. B. 2009. *Networks Evolving Step by Step: Statistical Analysis of Dyadic Event Data*, presented at the 2009 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, pp. 200–205.

Brunswicker, S., and Schecter, A. 2019. "Coherence or Flexibility? The Paradox of Change for Developers' Digital Innovation Trajectory on Open Platforms," *Research Policy*. (https://doi.org/10.1016/j.respol.2019.03.016).

Butts, C. T. 2008. "A Relational Event Framework for Social Action," *Sociological Methodology* (38:1), pp. 155–200.

Butts, C. T., and Marcum, C. S. 2017. "A Relational Event Approach to Modeling Behavioral Dynamics," in *Group Processes*, Springer International Publishing, pp. 51–92.

Chen, W., Wei, X., and Zhu, K. X. 2017. "Engaging Voluntary Contributions in Online Communities: A Hidden Markov Model," *MIS Quarterly* (42:1).

Chiu, C.-M., Hsu, M.-H., and Wang, E. T. 2006. "Understanding Knowledge Sharing in Virtual Communities: An Integration of Social Capital and Social Cognitive Theories," *Decision Support Systems* (42:3), pp. 1872–1888.

Choi, S. Y., Lee, H., and Yoo, Y. 2010. "The Impact of Information Technology and Transactive Memory Systems on Knowledge Sharing, Application, and Team Performance: A Field Study," *MIS Quarterly*, pp. 855–870.

Crawford, E., and LePine, J. 2013. "A Configural Theory of Team Processes: Accounting for the Structure of Taskwork and Teamwork," *Academy of Management Review* (38:1), pp. 32–48.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38.

DuBois, C., Butts, C. T., McFarland, D., and Smyth, P. 2013. "Hierarchical Models for Relational Event Sequences," *Journal of Mathematical Psychology* (57:6), pp. 297–309.

DuBois, C., Butts, C. T., and Smyth, P. 2013. *Stochastic Blockmodeling of Relational Event Dynamics*, in (Vol. 31), presented at the International Conference on Artificial Intelligence and Statistics, pp. 238–246.

DuBois, C., and Smyth, P. 2010. *Modeling Relational Events via Latent Classes*, presented at the Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 803–812.

Faraj, S., and Johnson, S. L. 2011. "Network Exchange Patterns in Online Communities," *Organization Science* (22:6), pp. 1464–1480.

Faraj, S., and Sproull, L. 2000. "Coordinating Expertise in Software Development Teams," *Management Science* (46:12), pp. 1554–1568. (https://doi.org/10.1287/mnsc.46.12.1554.12072).

Faraj, S., and Xiao, Y. 2006. "Coordination in Fast-Response Organizations," *Management Science* (52:8), pp. 1155–1169.
Granovetter, M. 1973. "The Strength of Weak Ties," *American Journal of Sociology* (78:6), pp. 1360–1380.
Granovetter, M. 1983. "The Strength of Weak Ties: A Network Theory Revisited," *Sociological Theory*, pp. 201–233.
Hansen, M. T. 2002. "Knowledge Networks: Explaining Effective Knowledge Sharing in Multiunit Companies," *Organization Science* (13:3), pp. 232–248.
Hansen, M. T., Mors, M. L., and Løvås, B. 2005. "Knowledge Sharing in Organizations: Multiple Networks, Multiple Phases," *Academy of Management Journal* (48:5), pp. 776–793.
Hedman, J., Srinivasan, N., and Lindgren, R. 2013. *Digital Traces of Information Systems: Sociomateriality Made Researchable*, presented at the Thirty Fourth International Conference on Information Systems.
Howison, J., Wiggins, A., and Crowston, K. 2011. "Validity Issues in the Use of Social Network Analysis with Digital Trace Data," *Journal of the Association for Information Systems; Atlanta* (12:12), pp. 767–797.
Kanawattanachai, P., and Yoo, Y. 2007. "The Impact of Knowledge Coordination on Virtual Team Performance over Time," *MIS Quarterly*, pp. 783–808.
Kankanhalli, A., Tan, B. C., and Wei, K.-K. 2005. "Contributing Knowledge to Electronic Knowledge Repositories: An Empirical Investigation," *MIS Quarterly*, pp. 113–143.
Karrer, B., and Newman, M. E. J. 2011. "Stochastic Blockmodels and Community Structure in Networks," *Physical Review E* (83:1), p. 016107. (https://doi.org/10.1103/PhysRevE.83.016107).
Kozlowski, S. W. J., and Klein, K. J. 2000. "A Multilevel Approach to Theory and Research in Organizations: Contextual, Temporal, and Emergent Processes.," in *Multilevel Theory, Research and Methods in Organizations: Foundations, Extensions, and New Directions*, K. J. Klein and S. W. J. Kozlowski (eds.), San Francisco, CA: Jossey-Bass, pp. 3–90.
Kudaravalli, S., and Faraj, S. 2008. "The Structure of Collaboration in Electronic Networks," *Journal of the Association for Information Systems* (9:10/11), pp. 706–726.
Kudaravalli, S., Faraj, S., and Johnson, S. L. 2017. "A Configural Approach to Coordinating Expertise in Software Development Teams," *MIS Quarterly* (41:1), pp. 43–64.
Lazer, D., Pentland, A. (Sandy), Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. 2009. "Life in the Network: The Coming Age of Computational Social Science," *Science (New York, N.Y.)* (323:5915), pp. 721–723. (https://doi.org/10.1126/science.1167742).
Leonardi, P. M. 2015. "Ambient Awareness and Knowledge Acquisition: Using Social Media to Learn" Who Knows What" and" Who Knows Whom"," *MIS Quarterly* (39:4), pp. 747–762.
Lerner, J., and Lomi, A. 2017. "The Third Man: Hierarchy Formation in Wikipedia," *Applied Network Science* (2:1), p. 24. (https://doi.org/10.1007/s41109-017-0043-2).
Majchrzak, A., Jarvenpaa, S. L., and Hollingshead, A. B. 2007. "Coordinating Expertise among Emergent Groups Responding to Disasters," *Organization Science* (18:1), pp. 147–161.
Majchrzak, A., and Malhotra, A. 2016. "Effect of Knowledge-Sharing Trajectories on Innovative Outcomes in Temporary Online Crowds," *Information Systems Research* (27:4), pp. 685–703.
Marks, M. A., Mathieu, J. E., and Zaccaro, S. J. 2001. "A Temporally Based Framework and Taxonomy of Team Processes," *Academy of Management Review* (26:3), pp. 356–376.
McFadden, D. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior," *Frontiers in Econometrics*, pp. 105–142.
Mesmer-Magnus, J. R., and DeChurch, L. A. 2009. "Information Sharing and Team Performance: A Meta-Analysis," *Journal of Applied Psychology* (94:2), p. 535.
Mesmer-Magnus, J. R., DeChurch, L. A., Jimenez-Rodriguez, M., Wildman, J., and Shuffler, M. 2011. "A Meta-Analytic Investigation of Virtuality and Information Sharing in Teams," *Organizational Behavior and Human Decision Processes* (115:2), pp. 214–225.
Quintane, E., and Carnabuci, G. 2016. "How Do Brokers Broker? Tertius Gaudens, Tertius Iungens, and the Temporality of Structural Holes," *Organization Science*.
Quintane, E., Conaldi, G., Tonellato, M., and Lomi, A. 2014. "Modeling Relational Events A Case Study on an Open Source Software Project," *Organizational Research Methods* (17:1), pp. 23–50.
Reagans, R., and McEvily, B. 2003. "Network Structure and Knowledge Transfer: The Effects of Cohesion and Range," *Administrative Science Quarterly* (48:2), pp. 240–267.

Robert, L. P., Dennis, A. R., and Ahuja, M. K. 2008. "Social Capital and Knowledge Integration in Digitally Enabled Teams," *Information Systems Research* (19:3), pp. 314–334.

Schecter, A., Pilny, A., Leung, A., Poole, M. S., and Contractor, N. 2018. "Step by Step: Capturing the Dynamics of Work Team Process through Relational Event Sequences," *Journal of Organizational Behavior*. (https://doi.org/10.1002/job.2247).

Shmueli, G., and Koppius, O. R. 2011. "Predictive Analytics in Information Systems Research," *MIS Quarterly* (35:3), pp. 553–572. (https://doi.org/10.2307/23042796).

Shore, J., Bernstein, E., and Lazer, D. 2015. "Facts and Figuring: An Experimental Investigation of Network Structure and Performance in Information and Solution Spaces," *Organization Science* (26:5), pp. 1432–1446.

Singh, P. V., Tan, Y., and Youn, N. 2011. "A Hidden Markov Model of Developer Learning Dynamics in Open Source Software Projects," *Information Systems Research* (22:4), pp. 790–807.

Sosa, M. E., Gargiulo, M., and Rowles, C. 2015. "Can Informal Communication Networks Disrupt Coordination in New Product Development Projects?," *Organization Science* (26:4), pp. 1059–1078.

Stadtfeld, C., and Block, P. 2017. "Interactions, Actors, and Time: Dynamic Network Actor Models for Relational Events," *Sociological Science* (4), pp. 318–352.

Tsai, W. 2002. "Social Structure of 'Coopetition' within a Multiunit Organization: Coordination, Competition, and Intraorganizational Knowledge Sharing," *Organization Science* (13:2), pp. 179–190.

Van de Ven, A. H., and Poole, M. S. 2005. "Alternative Approaches for Studying Organizational Change," *Organization Studies* (26:9), pp. 1377–1404.

Wasko, M. M., and Faraj, S. 2000. "'It Is What One Does': Why People Participate and Help Others in Electronic Communities of Practice," *The Journal of Strategic Information Systems* (9:2), pp. 155–173.

Wasko, M. M., and Faraj, S. 2005. "Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice," *MIS Quarterly*, pp. 35–57.