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

Aaron Schecter
University of Georgia, aschecter@uga.edu

Noshir Contractor
Northwestern University, nosh@Northwestern.edu

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

Completed Research Paper

Aaron Schecter
University of Georgia
Athens, GA
aschecter@uga.edu

Noshir Contractor
Northwestern University
Evanston, IL
nosh@northwestern.edu

Abstract

The widespread availability of digital trace data provides new opportunities for researchers to understand human behaviors at a large scale. Sequences of behavior, captured when individuals interface with an information system, can be analyzed to uncover behavioral trends and tendencies. Rather than assume homogeneity among actors, in this study we introduce a method for identifying 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. This study makes a contribution to the literature by introducing a novel methodology for analyzing sequences of digital traces. We apply our technique to data from a lab experiment featuring thirty twenty-person teams communicating over Skype.

Keywords: Digital traces; sequence analysis; unsupervised methods; social networks

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. We examine the information sharing behaviors of participants organized into a thirty team, 600 person lab sample. Our method identified two distinct patterns of behavior across the individuals in our sample. This study contributes 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

Relational event blockmodeling is an inductive method for determining different behavioral patterns among subsets of individuals. In general, we posit that individuals have some internal weighting scheme whereby different mechanisms carry different levels of influence on how they behave. Further, while each individual will vary in the criteria they apply, we anticipate that there will be a finite set of pattern classes (Chen et al., 2017; Singh et al., 2011). Essentially, there will be subsets of individuals in any group or "archetype" who exhibit similar sharing behaviors under comparable circumstances. These people are clustered into a single class defined by an aggregate set of decision-making criteria. Thus, we may predict information sharing behaviors in virtual teams by identifying a sufficient set of unique decision-making criteria that govern individual choices.

The relational event blockmodel builds on the standard relational event model (REM) (Butts 2008; Butts and Marcum 2017; Schecter et al. 2018) while introducing latent classes. Relational events are atomic units of interaction, containing information about a link and its timing (Butts 2008). Prior work has applied REM to the study of software development (Brunswicker and Schecter, 2019; Quintane et al., 2014), team communication (Schecter et al., 2018), email networks (Quintane and Carnabuci, 2016), and more. We proceed to detail the model construction.

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:

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

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):

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

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:

$$
\begin{aligned}
p\left(\beta_p \mid \mathcal{A}, z\right) &= \frac{p\left(\mathcal{A} \mid \beta_p, z\right) p\left(\beta_p, z\right)}{p\left(\mathcal{A}, z\right)} \\
&\propto p\left(\mathcal{A} \mid \beta_p, z\right) \\
&= \prod_{n=1}^{N} \lambda_{i_n j_n}(t_n)^{z_{i_n p}} \prod_{(i,j) \in \mathcal{R}} \exp\left(-(t_n - t_{n-1})\lambda_{ij}(t_m)\right) \\
&= \prod_{n=1}^{N} \exp\left(\beta_p' X_{i_n j_n}(t_n)\right)^{z_{i_n p}} \prod_{(i,j) \in \mathcal{R}} \exp\left(-(t_n - t_{n-1}) \exp\left(\beta_p' X_{i\,j}(t_n)\right)\right)
\end{aligned}
$$

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

Knowledge is a resource that is vital to organizational functioning and performance (Alavi & Leidner, 2001). To build organizational knowledge, the resources of individuals, e.g. information or expertise, must be efficiently combined. Essentially, organizations function as "open social systems that must process information…to accomplish internal activities, to coordinate diverse activities, and to interpret the external environment" (Daft & Lengel, 1986, p. 555). Following Nonaka (1994), we define information as "a flow of

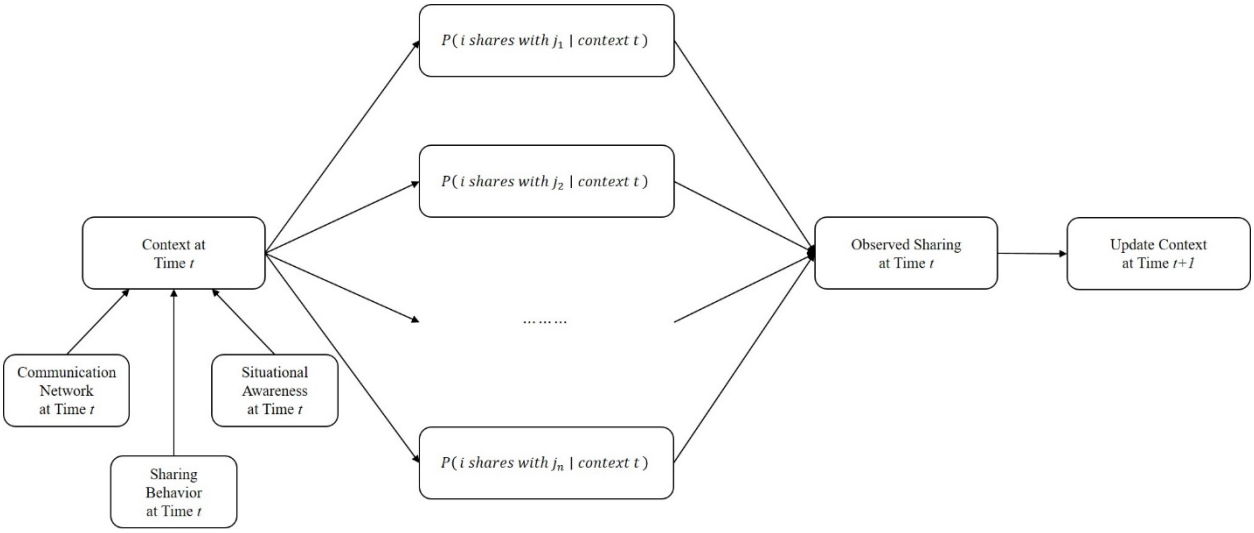
messages, and knowledge is created by the very flow of information" (p. 15). From this perspective, for an organization to generate the knowledge it requires there must be a sufficient flow of information among members of the group so that expertise can be effectively integrated (Robert, Dennis, & Ahuja, 2008). This process is particularly relevant in teams, which are often the building blocks of organizations. Information functions as a resource of the team, and teams convert their available resources into actions and or products (Alavi and Leidner, 2001). While individuals choose to share information for different reasons, we do not know a priori how each person will make their decisions. Instead, motivations for information sharing are only expressed through repeated actions. We focus on the context of ad hoc virtual teams because they lack significant transactive memory systems (Faraj and Xiao 2006; Majchrzak et al. 2007; Majchrzak and Malhotra 2016). In other words, members of these types of teams will identify pathways over time because they cannot rely upon meta-knowledge of who knows what or who knows who (Faraj and Sproull 2000).

We argue that focusing on individual events – i.e., instances of information being transferred – is necessary to understand how these emergent patterns form. We propose a process-oriented perspective where propensities to share information are formed dynamically as functions of both psychological constructs and explicit behavioral patterns (Poole, 2012; Van de Ven and Poole, 2005). A process-oriented approach treats information sharing as a series of events, i.e. the transfer of knowledge from one individual to another, that unfold over time (Quintane and Carnabuci, 2016). As such, individuals follow information sharing trajectories, where their propensities to act are continuously updated as new events occur and the context shifts.

Factors Influencing Information Sharing

We proceed to describe three factors that influence information sharing: the communication network, the sharing network, and individual situational awareness. These factors combine to form the internal motivation of each individual (Bock et al. 2005). We summarize the process of a sharing decision in Figure 1.

Figure 1. Architecture of sharing decision



Communication Network

We focus on four patterns from the informal communication network (Sosa et al., 2015) that we expect to affect sharing: 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. 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 be due to activity levels – i.e. an individual 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

In contrast to the informal communication network, the network created by sharing patterns describes technical communication amongst team members (Kudaravalli et al. 2017; Sosa et al. 2015). We focus on four measures that are parallel to the informal network: prior transference frequency, sharing reciprocity, total sharing activity, and sharing popularity. Prior transference frequency represents behavioral inertia, i.e., repeated sharing with the same individuals. Sharing reciprocity describes a tendency for individuals to share information with others who have previously sent them information (Faraj and Johnson 2011). Our third pattern, prior sharing activity, represents the tendency for an individual to maintain or expand their prior rate of sharing behavior. 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 a power law or centralization effect in the network.

Situational Awareness

We last consider measures related to the effects of time pressure and team progress (Marks et al. 2001). Specifically, we focus on the impact of time on decision making, specifically the time remaining to complete the task. We anticipate that an individual with strong situational awareness will change their behaviors to account for their time resources; by contrast, an individual with low awareness may not alter their behavior over time. The effect of time is also 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. Of course, other task-specific behaviors are possible. For instance, proportion of tasks completed could influence how individuals decide to share information. However, in this study we simply use time as a proxy for awareness.

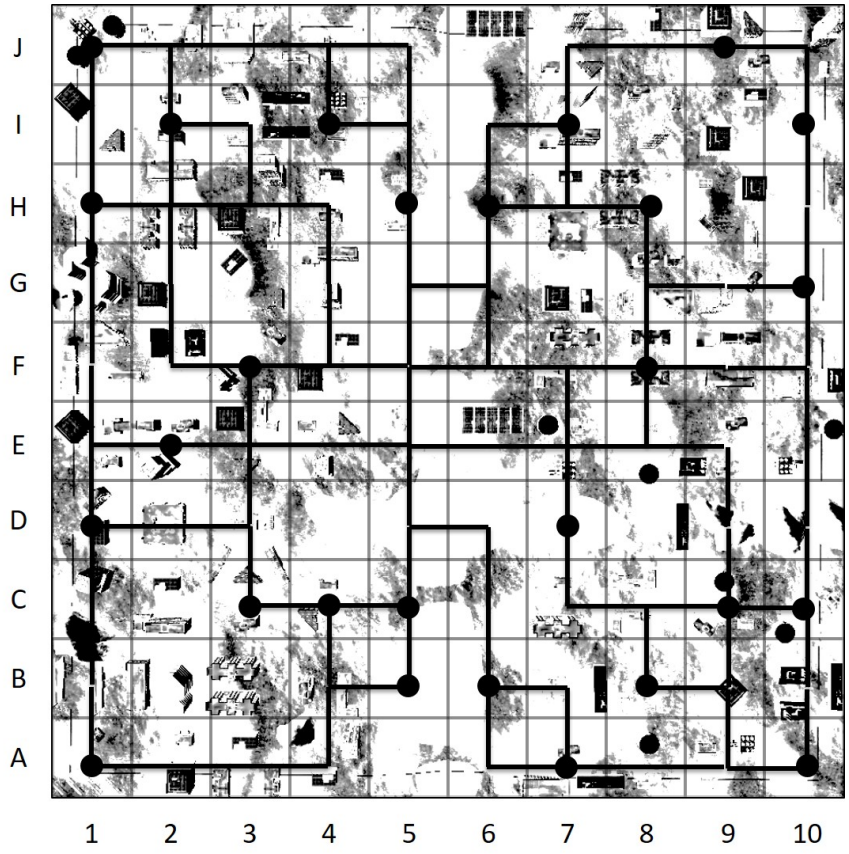
Methods

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. We chose a group size of twenty to ensure that 1) the teams were large enough to make information sharing difficult and 2) that the teams were small enough that information sharing could be accomplished in the limited time period. Each session in total covered a two-hour time period, during which participants took surveys, completed a practice mission, and then completed a performance mission. The practice period lasted twenty minutes, after which the participants were able to debrief. The performance mission took a total of forty minutes.

The task was a simulated military-style scenario in which participants cooperated to move a vehicle through an urban environment. In Figure 2, we show the game map and vehicle route. The objective was to move the vehicle as far as possible within the time limits, while avoiding damage. If the vehicle hit an obstacle, it had to stop until the participants cleared the threat out of the way. We demarcated the scenario map into four sections, and assigned each section to five participants. These five individuals formed a sub-group, and were responsible for all obstacles located within their quadrant.

Figure 2. Game map and route



We randomly assigned participants to one of the four sub-groups and to a role within that group. Each sub-group consisted of five individuals: a leader, a reconnaissance officer and a field specialist who work on counter-insurgency, and a reconnaissance officer and a field specialist who work on ordnance disposal. Each of the four teams had an appointed leader, and the leaders were charged with moving the convoy. The leaders had to agree on where and when to advance the convoy. The four non-leader team members were responsible for identifying and neutralizing threats. The reconnaissance officer was responsible for identifying potential threats and needed to communicate this information to the field specialist, who would then act on engaging and eliminating the threat.

Within each five-person unit, four individuals were provided with a unique set of coordinates. Each of these participants were given the locations of eight threats; two of those threats resided within their own quadrant, while six resided in other quadrants. The intelligence – totaling 144 unique threats – was provided randomly to participants on a sheet of paper at the beginning of the study.

Figure 3. 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.

By design, each participant had intelligence that was needed by *another player* in the game. While the participants were not directly incentivized to share information, the task could not be completed effectively if the coordinates were not used to clear targets on the game map. Thus, not sharing information would significantly reduce the likelihood of a successful mission outcome. In Figure 3, 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. All communication between participants occurred over Skype, including chats, calls, and video calls. Each player had an anonymous handle that indicated his or her 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. Information had to be shared through communication (verbal or non-verbal) using the designated channels, i.e., the Skype platform. The participants were free to share information at any time – there was no requirement to do so at the beginning or at any other specific instance.

From the Skype server we had access to a log of all messages sent across every session; our final data form is that of a transcript, with each row containing sender, receiver, time, and message. Manual coders identified pieces of intelligence in the individual messages, and marked those coordinates with a unique identifier. 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

In Table 1, we list our variables and relevant formulae.

Table 1. Operationalizations of focal measures

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	$x_9(i,j,t) = t$

We operationalized the mechanisms we described in the previous section by transforming the communication transcripts into sequences of relational events, taking the form (i, j, t, k) where k is an indicator for the intelligence contained in the message. The sequences were then compiled into two

temporal adjacency matrices, U and V, which we index by 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.

Analysis of Behaviors

To identify the unique patterns exhibited by the participants in our study, we applied the relational event blockmodel to the data from all 30 experimental sessions. We tested different numbers of archetypes as well as different subsets of parameters, starting from a single grouping. To determine the appropriate number of archetypes, we used the log-likelihood of the full models as well as the BIC to measure model fit. The best model should be one that achieves strong goodness-of-fit while also maintaining a degree of parsimony (i.e., not over-parameterizing).

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 -36,589, which was the lowest value among the candidate models. Thus, we conclude that there are two dominant archetypes in information sharing behavior. We present the comparison of model fit in Table 2.

Table 2. Model fit statistics for REM analysis

	P = 1		P = 2		P = 3	
	LL	BIC	LL	BIC	LL	BIC
Null Model	-4.2668	0.8534	-4.1142	0.823	-5.2684	1.0539
Network (N)	-4.0705	0.8144	-3.9048	0.7816	-4.8944	0.9798
Sharing (S)	-4.3412	0.8686	-4.2396	0.8485	-5.6601	1.1329
Awareness (A)	-4.061	0.8124	-4.0692	0.8142	-5.1505	1.0307
N + S	-3.83	0.7666	-3.6854	0.7382	-4.7263	0.9469
N + A	-3.8632	0.7731	-3.865	0.7739	-4.9281	0.9869
S + A	-3.8731	0.7751	-3.8506	0.771	-4.8748	0.9763
Full	-3.8189	0.7645	***-3.6589***	***0.7332***	-4.5046	0.903

Notes. Log-likelihood (LL) is reported as $(\ldots) \times 10^4$. *BIC is reported as* $(\ldots) \times 10^5$. *P is the number of distinct behavioral trends.*

To determine what factors make up these approaches, we examine the parameter values for this best model. The results are presented in Table 3. From Table 3 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.

Based on the computed ratios in Table 3, 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.

Table 3. Parameter estimates for two-archetype solution

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

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

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.

We also present in Table 4 demographic info for the two blocks we identified in order to rule out potential confounds (e.g., age or gender). We find that there are not significant differences across the two archetypes. Both groups have similar proportions of genders, are similar in age, and scored similarly on the ACT college entrance exam. On a 1-5 scale, both archetypes have average video game ability. Finally, we asked participants how familiar they were with the other individuals playing the game (1 = no familiarity, 2 = acquaintance, 3 = friend). In both groups, people reported knowing very few other members of their team before the mission. We thus conclude that these factors did not have undue influence on the behavioral differences observed.

Table 4. Demographic information for two-archetype solution

Measure	**Archetype 1**	**Archetype 2**
Percentage Male	54.21%	56.92%
Age	21.04	21.68
ACT Score	31.03	30.82
Video Game Experience	3.01	2.75
Participant Familiarity	1.17	1.18

Post Hoc Analysis of Performance

In order to determine post hoc if different sharing patterns were associated with variations in individual efficacy, we derived several measures of performance, including volume, accuracy, completion rate, and relative speed. Information volume was determined to be the number of messages sent by the individual that made reference to at least one piece of intelligence. Our second measure, information accuracy, was computed as the percentage of messages containing information that were directed to the appropriate person. We determined if the intelligence being shared was accurate by matching the unique identifier of the item being passed with the list of targets needed by the individual receiving the message. We only consider accuracy for individuals who shared at least one piece of information. Third, we computed a completion rate, which is the overall proportion of information shared that reached its target. Additionally, we computed a marginal completion rate that measures the proportion of messages that reached their target, conditional on them *not* being transmitted correctly the first time. Finally, we computed the average amount of time it took for information to reach its intended target. We measured this by denoting the time at which the focal individual possessed the intelligence, and then recording the time at which the information was first received by the correct target.

Given that we identified two distinct behavioral trends, we conduct a post hoc analysis to determine if performance outcomes vary across individuals in the different groups. In Table 5 we present the median values of our metrics across the two groups.

Table 5. Outcome measures for two-archetype solution

Measure	**Archetype 1**	**Archetype 2**
Messages	127	171
Information	6	6
Accuracy	0.25	0.23
Completion Rate	0.50	0.57
Relative Time	36.43	42.31

In table 5 we observe some differences between the two sets of participants. In archetype 1, individuals tend to send fewer messages overall. When they do share information, they tend to be *more accurate* on the first

transfer, but a *smaller proportion* of their information reaches its final target. However, for the information that does reach its target, it tends to arrive sooner when sent by members of archetype 1, relative to archetype 2.

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. With our method, it is possible to uncover the latent tendencies of each actor in a population. 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).

The findings of our empirical example 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). Further, we showed that the archetypes discovered are associated with distinct performance outcomes, despite having few underlying demographic differences. 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 bounded rational actors; of course, this implies that actors are able to accurately assess the state of the system concerning 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. 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. Finally, the use of time-dependent sequence statistics may lead to multicollinearity issues, depending on how the measures are constructed. To avoid these problems, we recommend testing models in a step-wise fashion in order to identify statistics which are unduly biasing the results.

With regard to the empirical example, the group size (20 people) and the laboratory environment may have had some influence on how well individuals could develop relationships and how easily information could be transferred. Future research in the field would help extend this work, as would a variety of group sizes and task types. Finally, other techniques such as text analysis (e.g., determining tone or affect) would provide additional nuance to our methods and provide a broader picture of the information sharing process. However, these techniques are 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).

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.

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.

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.

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.

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).

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.

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