Process Mining Organization Email Data and National Security Implications

John Bicknell and Werner Krebs

Abstract Many think of processes as sequential, deliberate activities which sustain businesses and government agencies; employees integrate themselves into defined organizational processes. From an ecosystem vantage, however, emergent processes exist and are discoverable. Emergent ecosystems form without human intention and may be especially influenceable. If emergent organizational processes–especially critical infrastructure processes–were explicit, they may be exploited. Tremendous intelligence is contained within semi-structured and unstructured organizational data sources. Properly analyzed, these data provide government and private organizations with actionable management and risk mitigation insights. Using explainable process technologies combined with natural language processing, a private critical infrastructure participant's organizational process model is discovered from semi-structured email data. Data derived from the process model are presented which elucidate internal operations and contribute to automated situational awareness of dynamically evolving events. National security implications and future research needs are described.

Keywords Process mining · Information warfare · Critical infrastructure · Organization modelling

1 Introduction

Many people think of processes as intentional groups of activities which serve profit maximizing businesses or mission driven government agencies. While this is true, information ecosystems and complex processes are also emergent and self-organizing with no human intention [1, 2]. Processes underlie all complex naturally occurring phenomena; indeed, the Earth is a process [3]. Process technologies

J. Bicknell (✉)
More Cowbell Unlimited, Inc., Lake Oswego, OR 97034, USA
e-mail: john@morecowbellunlimited.com

W. Krebs
Acculation, Inc., Los Angeles, CA 90036, USA
e-mail: wkrebs@acculation.com

D. Braha et al. (eds.), *Unifying Themes in Complex Systems X*, Springer Proceedings in Complexity, https://doi.org/10.1007/978-3-030-67318-5_15

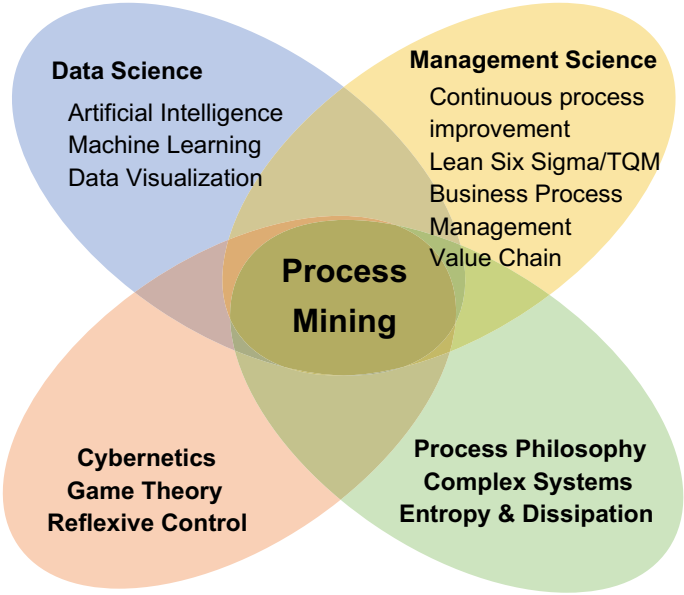

Fig. 1 Process mining is highly interdisciplinary. *Source* Authors

exist today which are an effective means to analyze time-domain relationships. As Fig. 1 suggests, process mining is inherently interdisciplinary–providing additive capability to data scientists, process improvement experts, operations researchers, organizational design practitioners, and national security analysts.

Process mining has its modern roots in Western Europe and was conceived originally in a business process improvement context. Data science and management science practitioners, Fig. 1, use process mining to speed the discovery of business process models while improving accuracy and preparing organizations for digital transformation [4]. Process mining is a highly versatile technique with vast utility beyond corporate process improvement. Poorly understood or opaque processes are a national security intelligence gap. If corporate, critical infrastructure, government, and societal processes were visible and explicit, opportunities for exploitation become available. Combined with other analytic disciplines such as game theory and complex systems theory (again, Fig. 1), process mining may inform critical infrastructure defense, information warfare (IW) vulnerability detection, network fragility analysis, geosurveillance, and many other analyses of national security importance.

Organizational process mining projects typically use structured enterprise resource planning system logs or other structured system data which record explicit activities, by specific employees, and at specific times. For example, a talent hiring process might include the following activities: Create Job Requisition; Post Job Announcement; Conduct Phone Screening; Conduct Phone Interview; Conduct Onsite Interview; Extend Offer; Accept Offer. These process steps are defined frequently

in structured system tables and are suitable for process mining with relatively little pre-processing.

In this study, we use a simple natural language processing (NLP) AI classifier to convert non-event email data into an IEEE standards compliant event log XES format [5], which facilitates process mining. Unstructured or semi-structured data must be pre-processed using an AI adapter. Previous email process mining efforts sought to understand business processes with a desire to improve business operations [6, 7] or to automate workflow discovery [8]; our research focuses on using business email data in national security contexts such as IW and critical infrastructure vulnerability detection.

Many defined business processes, as described earlier, produce Markov models which are absorbing. Our analysis is considerably different, although it uses the same IEEE data standard. We analyze an informal, emergent email communication process which is non-absorbing and ergodic; in other words, the process has no definitive beginning nor ending, and it is characterized by cybernetic looping behavior. Unlike previous email process mining efforts which required a tag in the subject line to identify a process activity [6] or which examined the entire email text for keywords [7], our intent is to show that an adversary may discern process information merely from untagged subject lines. Email subject lines often are considered less-sensitive than the full email and appear in SMTP log files.

Unstructured corporate data–including emails–contain significant business intelligence. It is therefore desirable to analyze email data in order to understand how a business or government agency might improve operations, identify vulnerabilities, and automate situational awareness of dynamically evolving events. In this paper, we present a process mining project which modelled an emergent corporate process using email data and process simulation technologies [9, 10]. We then discuss implications and suggest follow-on efforts.

1.1 Methodology

This study uses process mining, a nascent and powerful technique, to analyze the publicly available Enron email dataset [11].

The goal of process mining is to turn event data into insights and actions [12]. Process technologies elucidate ecosystem information flows, decision-making probabilities and temporal measures. Machine-readable process outputs and models of complex natural phenomena enable numerous applications. As the name implies, process mining AIs "mine" data and surface (e.g. Markov or Bayesian) models of decision-making processes from various input formats with no a priori knowledge. Process mining is also a human-understandable, human-verifiable, and human-explainable AI.

Structured or unstructured data which chronicle events are usable–cyber security logs, telemetry, information systems, multispectral imaging, social media data [13],

etc. Three pieces of information are needed to discover processes; additional features enrichen the analysis:

- Case ID: An identifier that represents a specific execution of a process.
- Activity: One of several steps performed within a process. For example, cyber exploit attempts or tweet hashtags.
- Time Stamp: This orders the activities within each case and enables sophisticated modeling.

Figure 2 contains a trivial example where an event log is converted into a temporal process model. Process cases 1 through 3 all contain the same activities, labeled "A" through "E." In Case 1, the activities happen in natural order. In Case 2, Activity "C" precedes "B." Finally, in Case 3, Activity "D" is repeated before concluding with Activity "E." The discovered process accounts for these process variations. Real world processes are significantly more complicated.

Process models are temporal representations of events organized by process case. The process mining software used in this analysis creates a machine-readable business process model notation (BPMN) diagram, calculates various Markov transition probability matrices, descriptive statistics associated with the process, capacity estimates for process activities, and distribution probabilities associated with activity state arrival data.

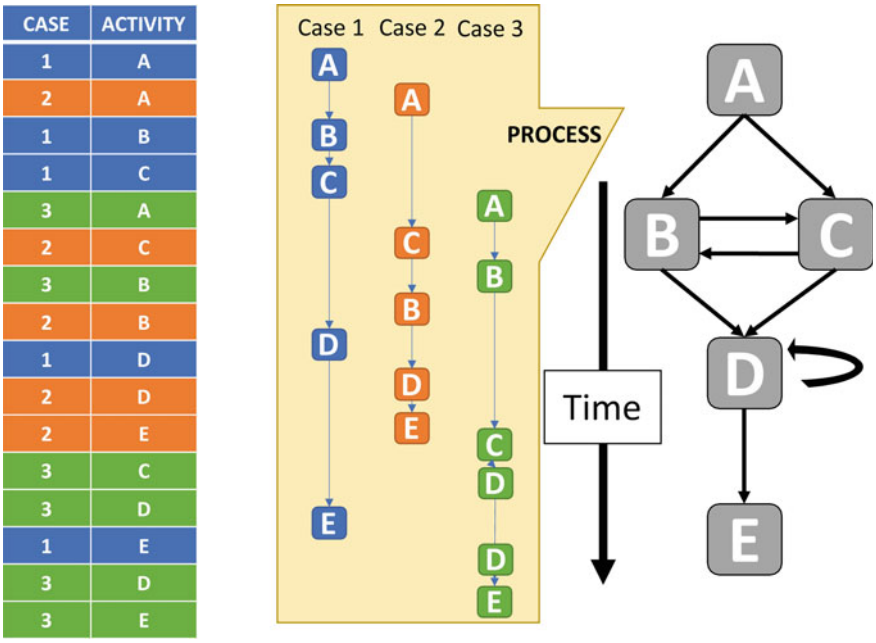

Fig. 2 Trivial process mining example

Fig. 3 Transition probability matrix state 1 → 2. *Source* Authors

State 1 \ State 2	A	B	C	D	E
A		0.67	0.33		
B			0.67	0.33	
C		0.33		0.67	
D				0.25	0.75
E					

The model presented in Fig. 2 may also be represented algebraically as a Markov transition probability matrix, Fig. 3. Information flows from beginning process activity states are expressed as probabilities. The beginning state is labelled "State 1" and is contained in the matrix rows; the ending state is labeled "State 2" represented by the matrix columns. Referencing the same simple event log contained in Fig. 2, information flows from Activity "A" to Activity "B" two-thirds of the time, and from Activity "A" to Activity "C" one-third of the time. When starting at Activity "D," there is a 25% probability that Activity "D" repeats and a 75% likelihood that Activity "E" follows Activity "D." The software used in this analysis uses a heatmap color coding scheme to visualize relative values and to suggest noteworthy findings.

Event log quality and richness varies greatly. The concept of a process sojourn is important. A sojourn comprises the path of a case between activities. Sojourn time may be measured from an event log and represents the entire activity wait and activity service times combined, Fig. 4. A sojourn for a case may proceed from Activity "A" to Activity "B", and so the sojourn time may be calculated as the Activity "B" End Timestamp minus the Activity "A" End Timestamp. If service start times are available, then explicit wait and services time calculations are possible. In a business context, sojourn measures describe waiting queues and service times for defined operational processes such as: parts manufacturing, loan processing, ecommerce webform user experience, or employee onboarding. Since this analysis uses email data, the only available sojourn measure is each email's send timestamp; therefore, it is impossible to calculate directly the sojourn wait and service times. So, we use the time between emails as a rough proxy for how much time is being spent on a

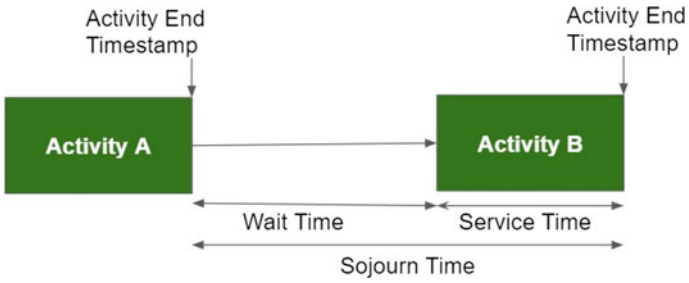

Fig. 4 Deconstructed process activity

task related to that email. This is an important consideration for estimating capacity of process activities, explained next.

Capacities for process activities may be estimated one of two ways. First, activity capacity may be estimated using simple queuing theory (SQT) comprising arrival rate and service rate estimates for activities derived from the event log; an alternative method for estimating capacity of an activity from the event log involves computing fractional spectral downshift in a frequency-domain representation of the event log. In this paper, we use the first method for estimating process activity capacities. Capacity is estimated using heuristics derived from SQT. Making a number of hand-waving assumptions, SQT-based heuristics look at the average number of tasks running in parallel together with sojourn times to estimate the number of process activities likely in a wait state at any given time; from this queue lengths, wait states, service times, and capacity may be estimated. We hypothesized that our very simple NLP email subject line event analysis, although noisy, would nevertheless serve as a rough proxy for the tasks employees were engaged in at any time. The resulting timestamps, in turn, when analyzed with SQT, in turn, provide a useful rough ranking of where employee bandwidth might be at a premium.

Descriptive statistics such as average, median, mode, skew, and kurtosis describe process sojourn times associated with state transitions. Probability distribution estimates with labels such as "Gaussian distribution," "extreme value distribution," "uniform distribution," "Poisson distribution," "platykurtic normal distribution," and so on are derived by applying heuristics to the statistical characteristics.

These process measures may be used as data-driven input parameters to simulate the organization. The analytic expression for the capacity parameter, defined above, is accurate under the assumption that there is only a single worker or single automated process processing events in the workflow. The expression, however, provides at least an estimate of capacity or utilization that is often directionally accurate as a simulation input parameter. Similarly, probability distribution descriptions are assumed to be approximately accurate.

Process mining and social network analysis (SNA) are conflated easily. While both techniques have a strong network component, process mining is inherently temporal, as explained earlier. Process models, therefore, may be analyzed using many of the same centrality, distance, and entropy measures associated with SNA. We describe future research possibilities which should explore these possibilities. A number of well-known visualizations and analyses [14–16] can be applied to the Markov models generated from process mining. Some of the more interesting outputs (Kemeny constant, relaxation time, mixing time, mean-time-of-first-passage, entropy, and eigenvalues, when present) have time units in terms of Markov steps. The process mining software used in this analysis generates "time-inhomogeneous" or "multiscale" Markov models with uneven time steps along with various time measures and estimates, including mean sojourn (transition times), and various techniques are known [17–19] to time-normalize or time-homogenize Markov matrices, after which Markov time step units are roughly proportional to real world time. Any method that generates XES process log data, including data from previous

process mining email efforts [6–8] can be subjected to our time-homogeneous, time-of-first-passage, capacity estimates, and resulting simulations.

Enron Corporation was an American energy, commodities, and services company based in Houston, Texas. It was founded in 1985. During the late 1990s, Enron was heralded as an American success story. Famously, Enron went bankrupt in late 2001 after accounting fraud and corporate corruption became public knowledge [20]. The Enron email dataset contains data from about 150 users, mostly senior management. The data contains a total of about 0.5 M messages and primarily spanned late 1998 through mid-2002 as Enron's collapse unfolded (although some spurious emails as well as incoming emails have timestamps outside this range). These data were originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation [11].

There are numerous ways to organize unstructured or semi-structured data for process mining. Different process event log configurations offer various temporal perspectives of information flow through an organization. For example, external email subject lines may be used as the process case in order to understand how external information flowing into an organization is handled. If department, team, or supervisory work group information is available, then these data provide aggregate views which offer comparisons between organizational sub-groups. For this analysis, the sender's and recipient's email addresses were used as CaseIDs (so that each retained email appeared as at least two events; bulk emails would appear as multiple events). A portion of the event log is contained in Fig. 5.

Process activities were derived using a simple, case-insensitive bag-of-words model [21]. We constructed a simple NLP classifier for the emails in Enron's dataset by looking at approximately the top-10, top-20, and top-40 non-stop-word email subject keywords. Emails were classified into unique activity categories (assumed to be approximately correlated with employee tasks) based on the first such top-10, top-20, or top-40 keywords that appear in the subject line. Emails with no non-stop-word subject lines, or emails which had subject lines not containing a top keyword, were dropped. An alternative, which was also attempted, is to classify such emails

	A	B	C
1	**Case ID**	**Start Time**	**Activity**
2	a..@enron.com	2001-04-19T15:32:27-07:00	meeting
3	a..@enron.com	2001-05-04T10:15:30-07:00	customer
4	a..@enron.com	2001-05-21T10:23:25-07:00	meeting
5	a..@enron.com	2001-05-29T15:30:24-07:00	SHOUT
6	a..@enron.com	2001-05-31T05:39:42-07:00	SHOUT
7	a..@enron.com	2001-05-31T09:03:24-07:00	SHOUT
8	a..@enron.com	2001-06-01T14:29:55-07:00	address
9	a..@enron.com	2001-06-05T13:15:46-07:00	meeting
10	a..@enron.com	2001-09-28T14:11:27-07:00	tickets

Fig. 5 Enron email event log excerpt

as ‘other’; this resulted in data which was deemed too noisy. In some of our data, we made one exception to this rule: noting that Enron frequently sent out emails with ALL-CAPS subject lines as a way of indicating an urgent email, if an email subject line was in ALL-CAPS (except for common control stop words like ‘Re:’), this was assumed to be a proxy for a crisis or urgent matter, and designated as belonging to the “SHOUT” process activity.

Process mining techniques include a number of filtering strategies for dealing with noisy datasets. Not all “SHOUT” Enron emails were crisis related; some were merely social events. Filtering process logs to retain only a top set of the most common activities is a heuristic referred to as an “activity filter” in the process mining literature. The subject line keyword in the last email sent or received by an employee with the top-keyword-set was assumed to be a proxy for that employee’s current activity in our very simple NLP model. The idea was to demonstrate that even very simple, easily automatable NLP analyses of unstructured corporate data could yield interesting process mining results. If an employee sent or received multiple emails containing that keyword in a row (even if this activity was interspersed with emails not containing subject lines of interest which were removed, as described previously) the employee was assumed to still be engaged with the subject keyword line activity, and the duplicate records were deleted.

The Enron dataset includes emails from outside the company, a small but significant percentage of corrupted emails, and emails from different email software systems (Enron transitioned off of Lotus Notes during this period). To ensure our dataset only looked at emails originating from within Enron and avoided bulk mail duplicates, we limited our data to emails found in ‘sent-mail’ folders, which may have been specific to one email system. By limiting to emails in employee ‘sent-mail’ folders, we initially processed only a subset of the 0.5 M Enron emails, looking at 37,921 emails, of which 37,830 were deemed non-corrupt, resulting (after ‘activity filtering’ on top keywords), in 4,994 events (which represented emails spanning multiple employee/CaseIDs as described above). Since ‘sent-mail’ may have been an artifact of one email system used by Enron, for comparison purposes we also did a second analysis of the entire Enron email archive (including emails sent from outside Enron). This comparison run looked at 517,401 emails, of which 495,466 were deemed non-corrupt. After “activity filtering” to limit events to approximately the top-10 subject line keywords as described above, this dataset had 63,306 events. Despite including emails from outside Enron and a different date range, this larger but presumably noisier dataset resulted in reasonably similar process mining statistics as the smaller ‘sent-mail’-only dataset.

2 Results

Derived process models combined with post hoc manual analysis revealed insights about Enron corporate culture and how management dealt in email with recurring corporate issues.

The Markov model contained in Fig. 6 represents Enron corporate decision-making or information flow through the lens of general email exchanges. Decision patterns emerge among Enron employees which are unintended, yet observable. The matrix includes a "re-work" perspective which is observable as the diagonal values. Re-work in this context means that an employee sent consecutive emails with a similar contextual subject, which is interpreted as engagement.

Emails with all capitalizations ("SHOUT") in the subject line take up a significant portion of employee email dialogue, as evidenced by the high transition probabilities going into the "SHOUT" state from other states. Moreover, there is a 55% probability that Enron employees remain in a "SHOUT" email re-work state. Responding to "meeting" emails is another large resource expenditure as evidenced with relatively high inflow probabilities and 47% probability of sending consecutive meeting-related emails (re-work). "Urgent," "Vacation," and "Please" are all activities with relatively little re-work.

The software used in this analysis auto-generates a narrative summary of the process:

> This process model was created by analyzing 172 cases. The average case time is 116.48 days, and the median case time is 93.07 days. The min case time is 0.00 s. The max case time is 431.99 days. It appears that the case durations are normal skewed to the right. There were 4994 total events processed. The activities which were estimated to take the longest amount of time are "address" at 7.45 days and "hi" at 7.26 days. There was an average of 29.03 events per case, and a median of 15.0 events per case. There were a max of 292 events per case and a min of 1. The most frequent activities are "SHOUT" at 1780 sojourns and "meeting" at 1265 sojourns. The process activities with the highest capacity estimates are "SHOUT" and "meeting."

	Activity Counts	hello	hi	lunch	SHOUT	address	conference	customer	game	holiday	imbalance	meeting	office	party	please	settlement	tickets	training	urgent	vacation
hello	81	.27	.04		.36	.01	.03		.01		.01	.11	.04	.04	.03		.01	.01	.01	.01
hi	58	.06	.31	.02	.31	.07	.02					.09		.02	.04	.02		.02		.02
lunch	69		.05	.32	.21	.02	.09	.03	.02	.02	.05	.08	.02	.03	.02		.02	.03		.03
SHOUT	1780	.02	.01	.01	.55	.02	.04	.02	.01	.01	.01	.16	.02	.02	.03	.01	.01	.01	.01	.02
address	115		.01	.03	.32	.20	.05	.05	.02			.18	.02	.05	.04		.02		.01	.01
conference	329		.01	.01	.24	.03	.28	.04	.01	.01	.01	.23	.02	.03	.03	.01	.02	.01	.01	.02
customer	134	.02		.01	.33	.05	.05	.18		.01	.01	.22	.01	.02	.03	.02	.01		.02	.02
game	70		.03		.22	.06	.06		.28	.01		.18	.03	.04	.04			.01		.03
holiday	54	.06			.20		.04			.26	.02	.20	.06	.04	.04		.02		.02	.04
imbalance	107				.22	.02	.11		.01		.42	.14	.01	.02	.02		.01			.02
meeting	1265	.01			.21	.01	.07	.02	.01	.01	.02	.47	.02	.01	.03	.02	.02	.02	.01	.02
office	114	.04		.01	.24	.01	.04	.02	.01	.03	.03	.23	.17	.01	.07	.04	.03	.01	.01	.02
party	168	.01	.01	.02	.29	.01	.02	.03	.01	.01		.10		.42	.04	.01	.02	.01		.01
please	195	.02	.03	.01	.31	.02	.05	.02	.01	.01	.03	.22	.01	.03	.17	.03	.01		.04	.02
settlement	94	.01	.01	.01	.23		.07	.02	.02		.01	.26	.01		.04	.26	.01		.01	.01
tickets	107	.02			.20	.01	.06	.01	.02	.01	.02	.26	.03	.01	.04	.01	.25	.01	.01	.05
training	78		.01	.03	.29			.03	.03		.03	.25			.03	.01	.05	.22	.03	
urgent	64		.02	.02	.32	.02	.02	.05				.24	.02	.02	.05	.02	.03	.02	.16	.03
vacation	111	.01		.03	.22	.03	.05	.06	.01	.01	.03	.23	.01	.05	.04	.03	.03		.01	.18

Transition probabilities less than .01 are not displayed.

Fig. 6 Enron email transition probability matrix with re-work

By removing re-work process events, transition probabilities for the "SHOUT" and "Meeting" activity states are more pronounced, Fig. 7. In fact, all of the matrix transition probabilities are slightly higher than the results contained in Fig. 6. This makes sense, as the process case activity event counts are reduced by the amount of re-work and the matrix row divisors are adjusted accordingly. Comparing the results from the model with re-work removed with the previous matrix (Fig. 7), we observe significant percent changes in information flows into "SHOUT" from the "Meeting" activity state, and vice versa. There is also a 22% change in probability that Enron employees send a "SHOUT" email following a "Party" email.

> This process model was created by analyzing 172 cases. The average case time is 116.48 days, and the median case time is 93.07 days. The min case time is 0.00 s. The max case time is 431.99 days. It appears that the case durations are normal skewed to the right. There were 2985 total events processed. The activities which were estimated to take the longest amount of time are "settlement" at 15.10 days and "game" at 12.93 days. There was an average of 29.03 events per case, and a median of 15.0 events per case. There were a max of 292 events per case and a min of 1. The most frequent events are "SHOUT" at 843 observations and "meeting" at 677 observations. The process activities with the highest capacity estimates are "meeting" and "SHOUT."

We hypothesized that our very simple NLP email subject line event analysis, although noisy, would nevertheless serve as a rough proxy for the tasks employees were engaged in at any time and provide a useful ranking of where employee bandwidth might be at a premium. Figure 8 compares the capacity estimates for Enron's email corporate model with activity event counts. "SHOUT" and "Meeting" have the largest capacity estimates at 0.92 and 0.85, respectively. The lowest capacity estimates include: "Urgent" at 0.23, "Training" at 0.24, "Settlement" at 0.26, and

	Activity Counts	hello	hi	lunch	SHOUT	address	conference	customer	game	holiday	imbalance	meeting	office	party	please	settlement	tickets	training	urgent	vacation
hello	61		.05		.49	.02	.04		.02		.02	.15	.05	.05	.04		.02	.02	.02	.02
hi	41	.08		.03	.46	.11	.03					.14		.03	.05	.03		.03		.03
lunch	48		.07		.31	.02	.13	.04	.02	.02	.07	.11	.02	.04	.02		.02	.04		.04
SHOUT	843	.04	.02	.03		.05	.10	.04	.02	.03	.02	.35	.04	.05	.07	.03	.03	.02	.02	.04
address	93		.01	.03	.40		.07	.06	.02			.23	.02	.07	.05		.02		.01	.01
conference	240		.01	.01	.32	.04		.06	.01	.01	.01	.31	.03	.04	.04	.01	.03	.02	.01	.02
customer	111	.03		.01	.41	.06	.06			.01	.01	.26	.01	.03	.04	.03	.01		.03	.02
game	51		.04		.31	.08	.08			.02		.24	.04	.06	.06			.02		.04
holiday	41	.08			.27		.05				.03	.27	.08	.05	.05		.03		.03	.05
imbalance	63				.38	.03	.20		.02			.25	.02	.03	.03		.02			.03
meeting	677	.01	.01	.01	.41	.02	.14	.04	.02	.01	.03		.05	.02	.06	.04	.04	.04	.02	.04
office	96	.04		.01	.29	.01	.04	.02	.01	.03	.03	.28		.01	.08	.04	.03	.01	.01	.02
party	100	.01	.01	.03	.51	.02	.03	.05	.01	.01		.17			.06	.01	.04	.02		.01
please	163	.02	.03	.01	.37	.02	.07	.02	.01	.01	.03	.26	.01	.04		.03	.01		.05	.02
settlement	70	.01	.01	.01	.31		.09	.03	.03		.01	.36	.01		.06		.01		.01	.01
tickets	81	.03			.27	.01	.08	.01	.03	.01	.03	.34	.04	.01	.05	.01		.01	.01	.06
training	61		.02	.03	.37			.03	.03		.03	.32			.03	.02	.07		.03	
urgent	54		.02	.02	.38	.02	.02	.06				.28	.02	.02	.06	.02	.04	.02		.04
vacation	91	.01		.03	.26	.03	.05	.08	.01	.01	.03	.29	.01	.05	.04	.03	.03		.01	

Transition probabilities less than .01 are not displayed.

Fig. 7 Enron s email transition probability matrix re-work removed

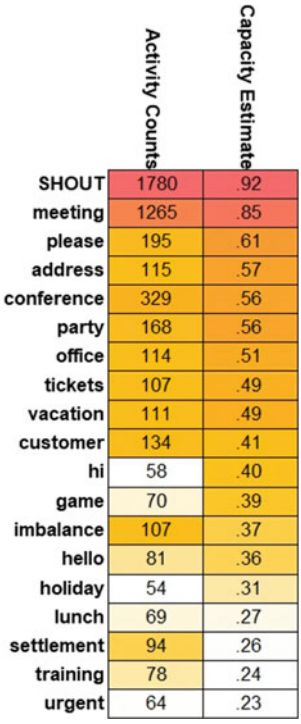

Fig. 8 Enron email process activity capacity estimates compared to event counts

"Lunch" at 0.27. Although there is a significant positive correlation between event counts and capacity estimates, $r = 0.83$, $n = 19$, $p < 0.001$, capacity estimates also looks at event timings and number of parallel tasks, so that "training" appears at significantly lower capacity than various social activities ("hi", "holiday", "lunch", "game") despite the former having higher or equal event counts (Enron was not noted for its compliance culture, so we might expect social activities to run at a higher capacity than "training" and "settlement", although it is even more surprising so many social-related keywords made it into the top keywords list. Given that "party" is above the median for both activity counts and capacity, it is perhaps not surprising that Fig. 7 shows a "SHOUT" response frequently following a "party" email, but not the reverse).

Skew, kurtosis, mean, variance and similar distribution statistics were applied together with simple heuristics to classify activity time interval probability distributions in the event logs, Fig. 9. For example, distributions with an absolute value or skew or excess kurtosis greater than 2 that are not classified as log normal are classified above as 'Extreme Value.' Similar heuristics classify distributions as likely 'Normal or Poisson', "Log normal', 'Leptokurtic normal', 'Platykurtic normal', 'Discrete', or 'Uniform.' These classifications can provide insights into the nature of processes and serve as data-driven simulation parameters. For example, processes with normally distributed sojourn times likely involve manual labor,

	hello	hi	lunch	SHOUT	address	conference	customer	game	holiday	imbalance	meeting	office	party	please	settlement	tickets	training	urgent	vacation
hello	4	4		4	7	1		7		7	4	6	6	1		7	7	7	7
hi	4	4	7	4	6	7					6		7	1	7		7		7
lunch		6	4	4	7	6	4	7	7	4	6	7	1	7		7	1		4
SHOUT	4	4	6	2	4	4	4	4	4	4	4	4	4	4	4	4	4	2	4
address		7	4	4	4	3	5	1			4	1	5	6		1		7	7
conference		1	4	4	4	4	5	4	6	4	4	6	5	4	6	4	6	6	5
customer	6		7	4	5	6	4		7	7	4	7	6	6	4	7		6	1
game		1		4	6	4		4	7		3	1	4	6			7		1
holiday	6			4		4			4	7	4	6	1	1		7		7	4
imbalance				4	1	4		7		4	4	7	1	1		7			1
meeting	4	6	4	4	4	4	4	4	4	4	2	4	6	4	4	4	4	6	4
office	6		7	4	7	4	1	7	6	6	4	4	7	4	4	4	7	7	1
party	7	7	6	4	1	6	5	7	7		4		4	3	7	4	1		7
please	6	4	7	4	6	4	4	7	7	4	4	1	4	4	5	7		4	4
settlement	7	7	7	4		4	1	1		7	4	7		4	4	7		7	7
tickets	1			5	7	6	7	1	7	1	4	4	7	6	7	4	7	7	6
training		7	1	4			1	4		1	4			1	7	6	4	1	
urgent		7	7	4	7	7	6				4	7	7	4	7	1	7	3	1
vacation	7		4	4	6	5	3	7	7	4	4	7	6	6	4	4		7	4

Legend:
1 = Discrete, 2 = Extreme value, 3 = Leptokurtic normal, 4 = Log normal, 5 = Normal, 6 = Platykurtic normal, 7 = Uniform

Fig. 9 Enron email probability distribution matrix

whereas processes with uniform or discrete time distributions may involve automated processes or indicate insufficient sample sizes. Processes near capacity often show log normal, leptokurtic or extreme value time distributions, likely due to employees utilizing triage to prioritize work queues. The Enron email process activities with the highest capacity estimates, "SHOUT" and "Meeting," are both estimated to have "Extreme Value" distributions.

"Normal modes" are a useful classifying characteristic in a number of dynamical systems [22, 23] and are associated with Markov processes when eigenvalues are present [23]. It is therefore reasonable to expect corporate processes with interesting eigenvectors to have normal modes, which might correspond to a steady-state-response to crisis (e.g. "SHOUT") and more relaxed ("game") corporate modes of operation. Eigenvalues have units of frequency, or inverse Markov step time, so a time-homogeneous Markov matrix may be more relevant to eigenvalue analyses. Not all corporate event traces will have interesting normal modes; typical corporate order flows have "absorbing" Markov Matrices without interesting loops and thus without interesting eigenvalues. However, our Enron-derived Markov chains have many looping processes, consequently they are ergodic and appear to have multiple interesting eigenvalues, which can be illustrated on a (time-inhomogeneous) eigen plot, Fig. 10.

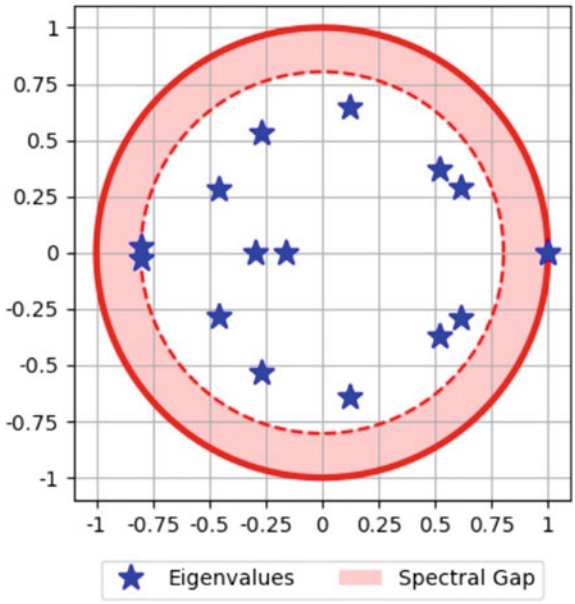

Fig. 10 Enron email process activity eigen plot

3 Discussion

Email process mining results using simple NLP AI model the flow of information and suggest Enron has a "meetings" culture and presumably has limited bandwidth for its total meetings and events. Based on our email subject line analysis, Enron meetings fall into different categories: trainings, client calls, conferences, networking or social event planning, and what appear to be crisis management and internal escalation meetings, all of which spawn each other at different rates. A crisis management meeting makes a training or conference call email, as well as future crisis management or internal escalation meetings more likely. Similarly, friendly meetings around networking or social events seem to make other social event-related meetings more likely. Thus, observation of the different categories of meetings impacts the probabilities of future instances of other meeting categories. A crisis-related meeting makes crisis-response and training meetings more likely as the company continues to respond to an ongoing crisis, whereas a social event meeting makes follow up social event meetings more likely as the company presumably relaxes and shifts "meeting bandwidth" towards these friendlier events.

Multiple process loops are present, resulting in a Markov matrix that is non-absorbing and ergodic with chain and topological entropies, relaxation time, and eigenvalues present. Since the dataset often shows switching between urgent ALL-CAPs ("SHOUT") subject lines and more relaxed social tasks, the Markov matrix presumably encodes data about Enron's response to various corporate crises and subsequent relaxation from those crises. Thus, eigenvectors, relaxation times, and

chain-entropy in a time-homogeneous Markov Matrix can potentially provide insights into information diffusion times and corporate response to external stimuli, one of many ways process-mining statistics may be used to infer corporate vulnerabilities or provide business consulting insights. Markov chain walk plots also provide a useful visualization of corporate task switching within our model.

Despite the noise in the underlying dataset and violations of single server SQT assumptions, the capacity algorithm nevertheless tends to identify work-related keywords such as "meeting", urgent "SHOUT"-related tasks, and other work-related keywords as among the top candidates for capacity drains, whereas the social-related keywords tend to receive lower capacity scores. Note that, because task parallelism is being factored in, this ranking is somewhat different than one achieved via simpler heuristics such as merely looking at activities with top counts or long-running tasks. Interestingly, Enron was not noted for its compliance culture, and at least in this particular time-slice set "training" has the second lowest capacity score of the top email subject line keywords, perhaps because it is lower priority than "settlement", "lunch," social introductions, or holiday parties. (In fairness to Enron, "training" would be a lower priority at many companies compared to "customer" if not "[social] party", and looking at other time slices suggest Enron's priorities in this regard may have changed a little over time. Nevertheless, given Enron's history, it is interesting that "training" scored quite low here despite considerable activity counts and sojourn times spent on the activity). Despite the low capacity score for "training", it is worth noting that many other keyword-activities in this same set would be lower ranked when merely looking at simpler criteria, such as raw activity counts or sojourn times. These results suggest that SQT analysis may add value as a useful ranking heuristic even on noisy data where SQT assumptions are violated. The software used in this analysis is also able to combine the Markov transition results with its activity timing estimates to automatically generate more accurate agent-based simulations free from the restrictions of SQT assumptions.

Conceptually, Enron employees are likely doing many tasks in parallel, but due to lack of data granularity and because this project was a first effort, we modeled Enron employees as working on one email task (process activity) at a time, even if there were minor distractions in between. Despite this simplification, our consistent approach appears to present a correct approximate picture of Enron corporate operations. Using a simple NLP classifier and simplifying assumptions about employee processing of work, this project constructed a process mining event log from semi-structured corporate email data. Post hoc analysis suggests promise for future research.

4 Implications and Future Work

In this section, we describe national security implications and future research considerations.

Government agencies and American corporations–especially critical infrastructure participants–must assume intelligent, persistent, and resourced adversaries

are gathering intelligence in order to launch insidious and potentially devastating IW attacks on the homeland [24–27]. Since the 1960s, Russia has enhanced IW with systematic psychological understandings of adversary reflexive processes and continues honing the technique. Known as reflexive control (RC), this highly analytical method has roots in cybernetics and game theory [24, 28, 29] and is a means of conveying to a partner or an opponent specially prepared information to incline him to voluntarily (or reflexively) make a predetermined decision desired by the initiator of the action [30].

IW enhanced with RC extends beyond human-to-human interactions; in fact, human-to-human IW RC is a minority use case in the Information Age. Today, entire societies and all public/private critical infrastructure participants must thwart human–machine and machine-machine IW attacks [25]. IW RC may be deployed with devastating effect against machines, information systems, and physical infrastructures [30]. Mapping of decision-making patterns, also known as "information cartography" [31], is a challenging but still achievable task. It is the knowledge of patterns within the decision-making process that allows an adversary to insert information into the process that would ultimately allow manipulation of the decision [32]. As such, all aspects of critical infrastructure ecosystems should be analyzed using a process-centric lens.

The informal, emergent corporate process contained in this study, for example, suggests ways information may be injected into the Enron ecosystem in order to disrupt operations, distract executives, or drain resources with frivolous activity. The processes discovered in this study represent probabilistic flows of information in the Enron corporate ecosystem. Future work should build upon this study and develop creative measures and simulations for elucidating process ecosystems further. Understanding ecosystem dynamics empirically answers operational questions of interest such as: how long will it take for specific information to reach a specific person, what is the organizational rhythm or tempo, how will the ecosystem react when information is injected at various entry points [33], how do adjacent ecosystems interact, and how long does it take for an ecosystem to recover? Measures such as the Kemeny constant, relaxation time, mixing time, mean-time-of-first-passage, entropy, and eigenvalues should all be considered. Time-normalized or time-homogeneous Markov matrices will likely provide additive ecosystem intelligence, as well.

Process models representing contextual temporal decision flows through organizations contain substantial intelligence from which organizational vulnerabilities may be inferred and mitigation strategies developed. Derived vulnerabilities may include reflexive reactions to deliberate or incidental stimuli (red team cyber-attack provocations, major financial upheavals, or weather disasters, for example), decision network fragilities/network points of failure, and diffusion of information both contextually and temporally. Future research may reveal important trends with profound national security implications. If enough organizations are studied and processes catalogued, generalized reflexive response trends to certain stimuli may emerge and be exploited in various ways such as using information as a maneuver element in a larger physical or cyber kill chain. New social engineering techniques may be developed as well which are, again, generalizable. Future research should include manipulating

Markov models in various ways to understand ecosystem perturbations empirically, simulating ecosystems to understand and predict information flows, investigating concepts like entropy warfare, entropy transfer, or directed entropy [34], developing cyber-attack/defense software which accepts machine-readable process model outputs [24, 25] to understand vulnerabilities through a new lens, and developing mitigation strategies which do not affect organizational operations.

As discussed, pre-processing tools are required for creating event logs; this is especially true for unstructured or semi-structured data. The simple bag-of-words NLP classification combined with human analyst adjustments methodology used in this project extracted keywords from email subject lines. New pre-processing AI adapters are necessary, however, in order to maximize future research. We hypothesize that useful process mining insights could be derived from the email-derived event logs using fully-automated contextual NLP AI adapters. If simple NLP techniques yield interesting results, imagine what can be done with more sophisticated natural language understanding (NLU) technologies, such as unsupervised document clustering, TensorFlow-based open source document NLU classification systems like Rasa, or proprietary systems like IBM Watson. For example, email bodies, email attachments, and entire corporate data systems are a much richer and untapped data source for process technologies. Process mining which combines customized contextual NLP from various perspectives (human resources, operations, finance, logistics, marketing, legal, etc.) enables comprehensive multi-dimensional organizational elucidation with greatly enhanced possibilities for whole-business optimization simulations and critical infrastructure vulnerability analyses.

5 Conclusion

This paper reported the findings of a novel application of organizational process mining. Enron corporate emails were analyzed after applying a simple NLP methodology to extract keywords from the email subject lines. The technique yielded interesting results and tells a logical post hoc story of internal operations which warrant further research to understand national security and organizational management implications.

References

1. Prigogine, I., Nicolis, G., Babloyantz, A.: Thermodynamics of evolution. Phys. Today **25**(11), 23–28 (1972). https://doi.org/10.1063/1.3071090
2. Hidalgo, C.: Why Information Grows: The Evolution of Order, from Atoms to Economies. Basic Books (2015)
3. Whitehead, A.N.: Process and Reality, 2nd edn. Free Press, New York (1979)
4. Bicknell, J.W.: Process mining technologies. ORMS Today **46**(5) (2019). https://doi.org/10.1287/orms.2019.05.01

5. IEEE: IEEE Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams. IEEE Std 1849-2016, pp. 1–50, Nov 2016. https://doi.org/10.1109/IEEESTD.2016.7740858
6. van der Aalst, W.M.P., Nikolov, A.: EMailAnalyzer : an e-mail mining plug-in for the ProM framework (2007)
7. Jlailaty, D., Grigori, D., Belhajjame, K.: A framework for mining process models from emails logs. ArXiv. abs/1609.06127 (2016)
8. Allard, T., Alvino, P., Shing, L., Wollaber, A., Yuen, J.: A dataset to facilitate automated workflow analysis. PLoS ONE **14**(2), e0211486 (2019). https://doi.org/10.1371/journal.pone.0211486
9. Bicknell, J.W., Krebs, W.G.: Methods and systems for estimating process capacity. United States 10,846,194, issued November 24, (2020)
10. Bicknell, J.W., Krebs, W.G.: Methods and systems for inferring behavior and vulnerabilities from process models. U.S. Patent Application No. 16/440,639. Washington, DC: U.S. Patent and Trademark Office
11. Enron Email Dataset, 18 May 2015. https://www.cs.cmu.edu/~./enron/. Accessed 12 Mar 2019
12. van der Aalst, W.M.P.: Process Mining: Data Science in Action, 2nd edn, 2016 edn. Springer, New York, NY (2016)
13. Bicknell, J.W., Krebs, W.G.: Detecting botnet signals using process mining (2019)
14. Khinchin, A.Y.: Mathematical Foundations of Information Theory, 1st Dover edn. Dover Publications, Mineola, NY (1957)
15. Kemeny, J.G., Snell, J.L.: Finite Markov Chains: With a New Appendix "Generalization of a Fundamental Matrix." Springer, New York (1976)
16. Belluzzo, T.: A framework for discrete-time Markov chains analysis (2019)
17. Feldman, J.F., Roberge, F.A.: The normalized transition matrix. A method for the measure of dependence between inter-spike intervals. Electroencephalogr. Clin. Neurophysiol. **30**(1), 87–90 (1971). https://doi.org/10.1016/0013-4694(71)90209-4
18. Hanks, E.M., Hooten, M.B., Alldredge, M.W.: Continuous-time discrete-space models for animal movement. Ann. Appl. Stat. **9**(1), 145–165 (2015). https://doi.org/10.1214/14-AOAS803
19. Bielecki, T.R., Cialenco, I., Gong, R., Huang, Y.: Wiener-Hopf factorization for time-inhomogeneous Markov chains and its application (2018). https://arxiv.org/abs/1801.05553. Accessed 02 Oct 2019 [Online]
20. McLean, B., Elkind, P.: The Smartest Guys in the Room: The Amazing Rise and Scandalous Fall of Enron, Reprint Portfolio Trade, New York (2004)
21. Bag-of-words model. Wikipedia, 07 Sep 2019. https://en.wikipedia.org/w/index.php?title=Bag-of-words_model&oldid=914476224. Accessed 18 Sep 2019 [Online]
22. Krebs, W.G., Alexandrov, V., Wilson, C.A., Echols, N., Yu, H., Gerstein, M.: Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic. Proteins Struct. Funct. Bioinforma. **48**(4), 682–695 (2002). https://doi.org/10.1002/prot.10168
23. Normal mode. Wikipedia, 08 Sep 2019. https://en.wikipedia.org/w/index.php?title=Normal_mode&oldid=914576448. Accessed 18 Sep 2019 [Online]
24. Bicknell, J.W., Krebs, W.G.: Process mining: the missing piece in information warfare. ResearchGate (2019). https://doi.org/10.13140/RG.2.2.23584.94722/1
25. Bicknell, J.W., Krebs, W.G.: FOCAL information warfare defense standard. ResearchGate, June 2019. https://doi.org/10.13140/RG.2.2.12672.07687.
26. Sviridova, A.: Vectors of the development of military strategy, 04 Mar 2019
27. Mueller, R.S.: Report on the investigation into Russian interference in the 2016 presidential election, Mar 2019. https://www.hsdl.org/?abstract&did=824221. Accessed 01 Oct 2019 [Online]
28. Chotikul, D.: The Soviet Theory of Reflexive Control in Historical and Psychocultural Perspective: Preliminary Study. Naval Postgraduate School, Monterey, California (1986)

29. Novikov, D.A., Chkhartishvili, A.G.: Reflexion and Control: Mathematical Models. CRC Press (2014)
30. Thomas, T.: Russia's reflexive control theory and the military. J. Slav. Mil. Stud. **17**(2), 237–256 (2004). https://doi.org/10.1080/13518040490450529
31. Waltzman, R.: SASC Testimony: The Weaponization of Information (2017)
32. Jaitner, M., Kantola, H.: Applying principles of reflexive control in information and cyber operations. J. Inf. Warf. **15**(4), 27–38 Fall (2016)
33. Ruocco, A., Buchheit, N., Ragsdale, D.: A combined offensive/defensive network model. In: 1st Annual IEEE Systems, Man, and Cybernetics Information Assurance Workshop, West Point, NY, June 2000, pp. 14–18. Accessed 22 Aug 2019 [Online]
34. Dobson, T.K.: Entropy and self-organization—an open system approach to the origins of homeland security threats. Thesis, Naval Postgraduate School, Monterey, California (2015)