Balancing Efficiency and Risk in Procure to Pay: Safely Realizing Cost Savings Using Process Mining Techniques

Adrien Porter, David Masse, Nuss Visatemongkolchai, Jithendra Seneviratne, Tanvee Deokule, Nicholas Hartman

CKM Analytix, 200 West 41st Street, New York, NY 10036, USA
{aporter, dmasse, nvisatemongkolchai, jseneviratne, tdeokule, nhartman}@ckmanalytix.com

Abstract. For the 2019 BPI Challenge, we use process mining techniques to explore and analyze procure to pay event logs from a large multinational paints and coatings company. Suboptimal procurement processes can lead to increased costs and operational risks for businesses. Within the company's data, we identify a range of opportunities for process optimization, including:

1. Invoices are frequently paid at only fixed pre-defined intervals. These intervals, which vary by vendor, often introduce long delays in payments while the process awaits the next available payment window. We model that clearing invoices more dynamically could better optimize cash flow management and allow for negotiation of discounts based on faster payments.

2. We apply algorithms to scan for noncompliance against published process requirements and recommend user re-training and/or system adjustment for users or vendors associated with concentrations of such behaviors.

3. We analyze the existing use of automation throughout the process, which identifies several activities handled via a mix of human resources and automated systems. We model the impact of a potential increase in the saturation of automation and make recommendations on where such automation is best targeted.

4. We rank vendors using a custom two-dimensional complexity metric that identifies which vendors most commonly cause common process inefficiencies. By flagging the least consistent and most time-consuming vendor-specific processes, future process efficiency efforts can be better targeted.

5. A social network analysis reveals that certain users perform more than one essential cross-checking step necessary for payment processing, which may raise the potential for fraud or errors by limiting key checks and balances. We also identify vendors and categories of goods/services that are particularly prone to such behaviors.

6. We identify vendors and categories of goods/services with frequent change events (e.g. "Change Quantity" or "Change Approval for Purchase Or-

der"). Modeling of the impact of these activities identifies that such rework introduces significant delays within the overall process and increases the amount of human labor required to complete the process.

By leveraging ongoing advanced process mining to monitor the impact of actions against the above opportunities, the subject company will be able to realize material cost savings and efficiency gains while also better monitoring and managing risks associated with the procure to pay process.

Keywords: Process Mining, Process Discovery, BPIC 2019, Process Improvement, Event Logs, Conformance Checking, Social Network Analysis, Automation, Procurement, SAP, Materials Management, Machine Learning

1 Introduction

Procurement is a critical business activity for any enterprise. Businesses commonly utilize purchase orders to acquire the goods and services necessary for their operations. Modern businesses make use of enterprise resource planning (ERP) systems to digitally track the progress of each order. These systems produce valuable data that can be mined to answer a multitude of questions about the functioning of the procurement process. In this report, we explore and analyze the BPI Challenge 2019 dataset to optimize the process, identify and manage risks and provide additional leverage for future pricing negotiations. We perform analyses on compliance, automation, throughput, payment times, process complexity, and social networks.

2 Overview of the Data

2.1 The Data

The BPI Challenge 2019 dataset comprises just over one year's worth of purchase order data from a large multinational paints and coatings company. The raw dataset contains 1,595,923 events distributed across 251,734 cases [1]. Cases are defined as a combination of a purchase order and purchase item. Each time-stamped event contains one of 42 activities such as "Create Purchase Order Item." In addition, each event has an associated set of informational attributes which includes the vendor, the value, the categorizations of the purchased item, the ERP system user, and the type of purchase order. Based on the terminology present in the data, we infer that the system in use is a commercial product made by SAP.

The dataset contains some cases with event timestamps that fall outside of the date range specified in the challenge instructions. We apply a date filter to only allow cases that start and end between the beginning of 2018 (2018-01-01 00:00:00) and the publication date of the dataset for the competition (2019-01-27 23:59:59). This reduces the dataset to 1,587,802 events across 251,463 cases. 99% of the data is retained after this filter is applied. We present detailed descriptions, statistics, and observations about the time-filtered data in the appendix.

The data was processed and analyzed using a combination of Fluxicon Discovery, ProM, and a process-mining tool internally developed at CKM Analytix, as well as custom Python analytics code.

Several sources of the event log were published for the competition. We noticed a time-zone discrepancy (events shifted by five or six hours) between the CSV data from the BPI Challenge web page and the .dsc file provided by Fluxicon. All analyses discussed in this report are based on the raw CSV data.

Table 1 shows summary statistics around case completion. Because the dataset is a slice in time, it is important to distinguish between cases that can be considered complete (i.e. contain a "Clear Invoice" event or "Record Goods Receipt" in Consignment cases) and cases that may have been in process when the sample was taken. Every case includes exactly one "Create Purchase Order Item" event, so the beginning of each case is present (i.e. no case is included in the log that started before the timeframe of the log and continues in the log). Time-related aggregates (e.g. median case duration) do not apply to incomplete cases as an incomplete case may have just started but could last any number of days into the future.

Table 1. Case Completion Statistics

	Count	% by Count	Value (EUR millions)[1]	% by Value	Mean Value (EUR)	Median Value
Completed cases	196,881	73.2%	711.6	78.3%	3,615	491
Non-completed cases	54,582	26.8%	260.1	21.7%	4,766	565
Total	**251,463**	**100%**	**971.8**	**100%**	**3,864**	**508**

2.2 Description of the Four Archetypal Processes

An important aspect of the data in this challenge is the existence of sub-processes explicitly specified in the challenge statement and codified in the data based on certain attributes (see "case Item Category" in the appendix). Below we present an overview of each process and summarize associated key statistics in **Table 2**.

[1] Sum of values recorded for "Create Purchase Order Item," which occurs exactly once per case.

Table 2. Summary Statistics for the four archetypal processes

	3-Way After	3-Way Before	2-Way	Consignment
Events	312,554[2]	1,233,410	5,758	36,080
Cases	15,129	220,810	1,027	14,497
Complete Cases[3]	9,624	173,503	289	13,465
Unique Activities	38	39	11	15
Unique Process Variants[4]	5,297	8,591	144	301
Median Complete Case Duration (days)	80	77	7	20
Mean Complete Case Duration (days)	89	81	20	24
Median Case Value (EUR)	402	594	6,174	0

The process maps below, for clarity, show only the most common and essential steps for compliance. The numbers between the straight arrows at the left show median position ("rank") within the cases for each activity. The numbers next to the curved arrows show the number of transitions from one activity to another in the event log (thickness of the arrows is scaled by this figure).

3-way Match, Invoice After Goods. We refer to this process as "3-way-after" throughout the rest of the report. Invoice receipts should be entered only after goods are received (activity "Record Goods Receipt") and are matched against the goods receipt and PO creation. We include the Service subprocess as part of "3-way-after" in the summary statistics but isolate and analyze it in compliance section of the report.

---

2 This figure includes the Service sub-process, which comprises 261,016 events and 5,800 cases within the main 3-way after process.

3 Complete Cases: Cases are considered complete if they include both "Create Purchase Order Item" and "Clear Invoice" activities ("Record Goods Receipt" for Consignment). Complete case durations are based on the time elapsed between the first instance of a "Create Purchase Order Item" activity and the last instance of a "Clear Invoice" activity.

4 Calculation of variants: we group together otherwise similar cases with different counts of repeated events at the same moment. For instance, a case with event sequence A → B (recorded 3 times at the exact same timestamp) → C would be considered the same variant as a case with sequence A → B (recorded 5 times) → C.

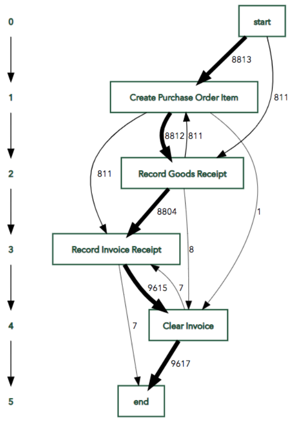

Fig. 1. First cycle (first instance of activities up to the first “Clear Invoice”) of complete cases

3-way Match, Invoice Before Goods. We refer to this process as “3-way-before” throughout the rest of the report. In this process, an invoice receipt may be entered prior to the entry of a goods receipt, but any payment is blocked until the goods receipt is entered and matched against the invoice received and PO created. We define “cycle” as the series of activities leading up to an individual instance of “Clear Invoice,” i.e. a payment to a vendor (see compliance section 3.2 for a description of how corresponding events are identified). For 91% of the cases in this group, the goods receipt occurs before or at the same time as the invoice receipt (as in “3-way match, invoice after goods”), at least in the first cycle. For compliance, these cases do not require “Remove Payment Block,” but 21% of the time, this extra step is performed, increasing the median time between “Create PO Item” to “Clear Invoice” from 72 to 91 days.

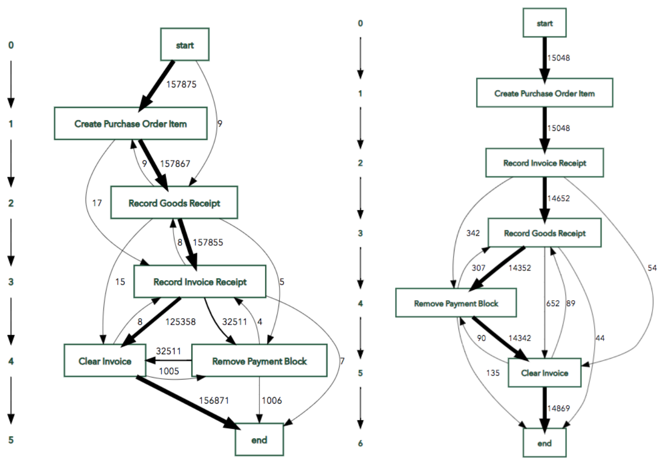

Fig. 2. First cycle of complete cases where “Record Goods Receipt” happens before or simultaneously with “Record Invoice Receipt” (left) and where “Record Goods Receipt” happens after “Record Invoice Receipt (right)

Approximately 95% of cases where “Record Invoice Receipt” happens before “Record Goods Receipt” do have “Remove Payment Block” as required for compliance and take a median of 73 days to complete.

2-way Match. We refer to this process as “2-way” throughout the rest of the report. Invoices received are simply matched to the initial purchase-order value.

Consignment. We refer to this process as “Consignment” throughout the rest of the report. Goods receipts are matched against initial purchase-order value only as there are no invoices associated with this sub-process.

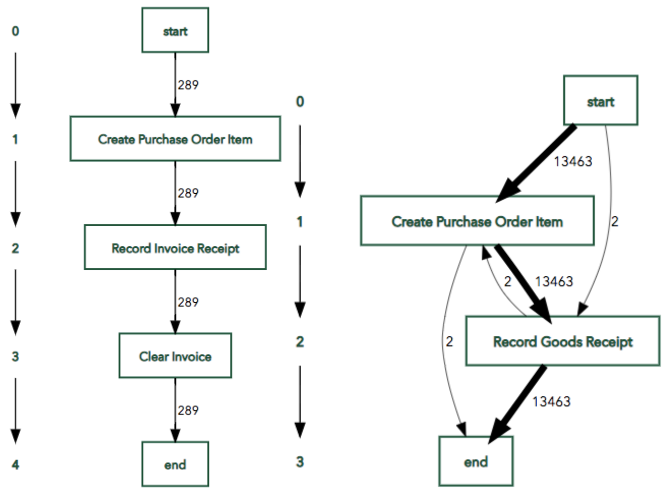

Fig. 3. First cycle of complete cases for 2-way (left) and Consignment (right)

3 Compliance Checking

3.1 Key Findings and Recommendations

Compliance Analysis. We deem a process to be compliant if it follows a set chronological sequence of events in conjunction with matching EUR amounts. These sequences are determined for each process type, as defined in the challenge statement. Cases belonging to the “Service” Item Type (subprocess of 3-way-after) exhibit low compliance rates (43.7% by number) compared to other processes (>80%). We recommend adjustments to the manner in which the system is utilized, either by reprogramming batch resources and/or retraining human users in order to maintain better adherence to standard PO process guidelines, thereby reducing financial and operational risk.

Least Compliant Vendors. Our analysis reveals a set of high-value vendors associated with low compliance rates. These processes should be scrutinized, as they could present a substantial financial risk to the company. The highest-value vendors that comply less than 90% of the time are ID_0183, ID_0479, ID_0234, ID_1023 and ID_0404.

3.2 Compliance Analysis

In order to determine which cases have been handled properly, we must define a compliance pattern for each sub-process in the event log. This pattern is a sequence of activities that must occur in time order and with matching EUR amounts for a case to be called compliant. “Create Purchase Order Item” occurs once and only once per case throughout the log, but other activities can occur zero, one or more times per case. For a case to be compliant, we require a correct sequence of matching events to precede every instance of “Clear Invoice” in a case, i.e. every cycle must be compliant. The one “Create Purchase Order Item” event is used for all cycles, but to match up other events, we number each instance of a particular activity with its time order, i.e. “Record Goods Receipt 1” then “Record Goods Receipt 2” and so on. Thus “Record Goods Receipt 2” is checked against “Record Invoice Receipt 2” and possibly “Remove Payment Block 2” for proper sequencing and EUR value. We cannot judge whether an incomplete case will ultimately reach a compliant outcome, so we only apply our compliance test to completed cases.

Compliance by Process. Below we present the patterns used for each of the four main sub-processes. Events with a EUR value that does not match the EUR value of the item at PO creation are addressed separately. Simultaneous events are considered “in compliant order” regardless of their positions in the event log. A summary is shown in **Table 3**.

Three-way matching, invoice after goods. Each instance of “Clear Invoice” must be preceded by the following sequence in order: “Create Purchase Order Item,” “Record Goods Receipt,” “Record Invoice Receipt.” 3-way-after cases have been further split into cases with Item Type Service and cases with non-Service Item Types (see discussion below).

Three-way matching, invoice before goods. Cases that are compliant using the pattern for “3-way-before” are deemed compliant. In addition, “Record Invoice Receipt” is allowed to occur before “Record Goods Receipt”; but when this happens, compliant cases need to have a “Remove Payment Block” event after “Record Goods Receipt” and before “Clear Invoice.” This is to ensure the invoice is not paid before the goods ordered are confirmed as received.

Two-way matching. Each instance of “Clear Invoice” must be preceded by the following sequence in order: “Create Purchase Order Item,” “Record Invoice Receipt.”

Consignment. “Create Purchase Order Item” followed by at least one instance of “Record Goods Receipt.”

Table 3. Compliance for the four archetypal processes and Service sub-process

	3-Way After	Non-Service	Service	3-Way Before	2-Way	Con-sign.
Compliance by number	80.5%	93.8%	43.7%	96.8%	100%	100%
Compliance by EUR value	80.2%	83.1%	79.4%	95.4%	100%	n.a.

Analysis. To address the markedly low Service compliance numbers, our analysis uncovers two primary driving forces behind noncompliance for such processes.

Mismatched EUR values. We note that 1.1% of completed cases have at least one event with a EUR value that does not match the EUR value of the PO item at creation. Among cases with Item Type Service, this figure jumps to 53%. The non-matching EUR values for “Record Goods Receipt” and “Record Invoice Receipt” are often near-multiples of the expected value. We assume that the fact that the multiples are not exact arises from rounding errors during the anonymization of the data as discussed on the challenge description page [2], or from tax, shipping or other add-on charges.

Since we do not have the quantity for each PO item, we could not determine with certainty which cases were handled properly. These non-matching values may indicate 1) confusion around when to use unit prices and when to use aggregate values in the system and/or 2) a problem with the way service entry sheets, upon approval, are translated into goods receipts (typically done by batch_06).

We noticed that the median time between PO item creation, service entry sheets and goods receipts for Service cases is zero, so this may be a case of automating a process that should not occur. The system should check that the number of these events makes sense and prohibit extraneous events or those with incorrect EUR values.

Lack of standardized usage of the system across Service processes. We grouped together similar processes based on event sequences using K-modes clustering. This method is chosen due to its performance on categorical data and scalability for large datasets. Four clusters emerged as a model that best describes the data as determined by the silhouette method, which finds the optimal separation of data based on their similarities and differences.

Of these four groups, one appears to be the standard process, with core activities with no repetition (881 cases, of which 67.2% compliant). The remaining three groups, however, see significantly lower compliance rates with an average of 27.3% across a combined 1,677 cases. Processes within these groups see distinct deviation from the standard process, such as with ‘Record Goods Receipt’ and ‘Record Service Entry Sheet’ appearing with N repetitions simultaneously. EUR value recorded at invoice receipt may also be an N multiple of the value at purchase order creation.

These different groupings highlight potentially inconsistent usage of the system by Service vendors. We therefore recommend standardizing procurement system practices where possible to simplify compliance detection and reduce risk.

Data Exploration using a Random Forest Classifier. In parallel with compliance analyses discussed above, a random forest classifier was applied to the dataset to help uncover patterns behind compliance. We utilized the classifier as a data exploration tool find meaningful data segmentations (e.g. vendors, spend areas) given compliance outcomes by examining characteristics of a case or event that most heavily influence the decision tree. While direct results from this analysis were omitted to respect the limited length of this report, we note that the methodology has been an integral tool in extracting insights from this dataset. An elaboration of how the model works can be found in the appendix.

3.3 Least-Compliant Vendors

Having established a framework for compliance, we are able to identify which vendors are associated with a high degree of non-compliant processes. We look at 3-way-before and 3-way-after cases together since vendors are present across different subprocesses, while setting aside 2-way and Consignment cases since they are 100% compliant. The bubble plot in **Fig. 4** further highlights the low compliance of Service processes, as discussed in the prior section. Additionally, we notice a higher degree of compliance among higher volume vendors. Nonetheless, we also see a certain number of vendors that stand out based on their lower levels of compliance. The figure also shows that the number of cases does not necessarily correlate with EUR value of the vendor's cases (sum of PO item creation event values, shown as circle size below). We therefore extracted the vendors who are responsible for the highest EUR value in items but were compliant for under 90% of their cases. We chose these vendors due to the fact that non-compliance for high value purchase orders presents a substantial risk to the company.

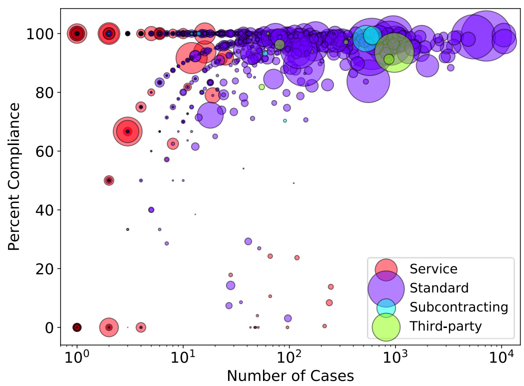

Fig. 4. Vendor compliance with number of cases. Bubble size indicates cumulative value of items from a vendor. Note the logarithmic scale on the horizontal axis.

We present a set of vendors that warrant further investigation in **Table 4**.

Table 4. Highest-value vendors that comply less than 90% of the time

Vendor	Number of Cases	Compli-ance %	EUR (millions)	Doc. Type	Item Type	Spend Class	Spend Area
ID_0183	555	83.8	17.0	Standard PO	Standard	PR	Latex & Monomers
ID_0479	137	89.1	15.6	Standard PO	Standard	PR	Titanium Dioxides
ID_0234	3	66.7	7.8	Framework order	Service	NPR	Logistics
ID_1023	18	72.2	6.3	Standard PO	Standard	PR	Titanium Dioxides
ID_0404	2001	89.1	4.7	Standard PO	Standard	NPR	Sales

Case Study: vendorID_0183. A deeper look at this particular vendor reveals that non-compliance tends to occur as a result of non-compliant activity occurring after a compliant first cycle (from "Create Purchase Order Item" to "Clear Invoice"). We identify a number of cases where a "Cancel Invoice Receipt" event occurs over 120 days after the initial "Clear Invoice" event, which would conclude a compliant cycle. Following "Cancel Invoice Receipt," we see a new "Record Invoice Receipt", and a new "Clear Invoice", all happening within a few minutes of each other. We suspect that this sequence is a result of some kind of necessary price/quantity change that could be benign. However, we highlight the fact that a "Clear Invoice" event has already occurred (and presumably funds have been transferred to the vendor), and that this post-completed transaction work and payment do not appear to follow a robust process. The system records no information about what could be causing this, making audit difficult. Either the user has not been educated on how to perform these types of actions, or the system is not set up to allow a process which should be able to take place.

4 Automation

4.1 Key Findings and Recommendations

Automation Opportunities. 49% of events in the log show potential for automation, representing an opportunity for the client to reduce payroll costs. Analysis reveals that automated activities are presently concentrated in certain spend areas and vendors. We recommend that the company begin investigating these processes to further automate its workflow.

4.2 Automation Analysis

We explored activities by resource type: human users or batch systems, which are described by the challenge as automated processes. Several activities that could be automatically executed by a batch resource were also completed by human users (**Fig. 5**).

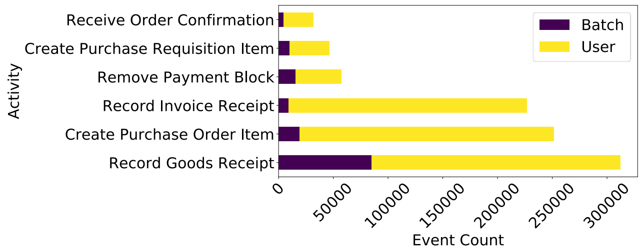

Fig. 5. Top six automatable events by count.

Human-user-executed instances of these the top six events by count represent 49% of events in the dataset[5]. Without additional information from the client at this stage, we assume that this represents a significant opportunity to further automate the workflow in order to reduce costs. Why were such events logged as done manually? Could they have been instead carried out by batch processes?

A key assumption in this analysis is that 'user' and 'batch' resources are synonymous with manual and automated. Degrees of automation differ widely across 1,961 vendors. We noticed instances of events that were carried out by users but occurred on a very consistent schedule (e.g. every 12 am on Sunday). These occurrences are relatively rare.

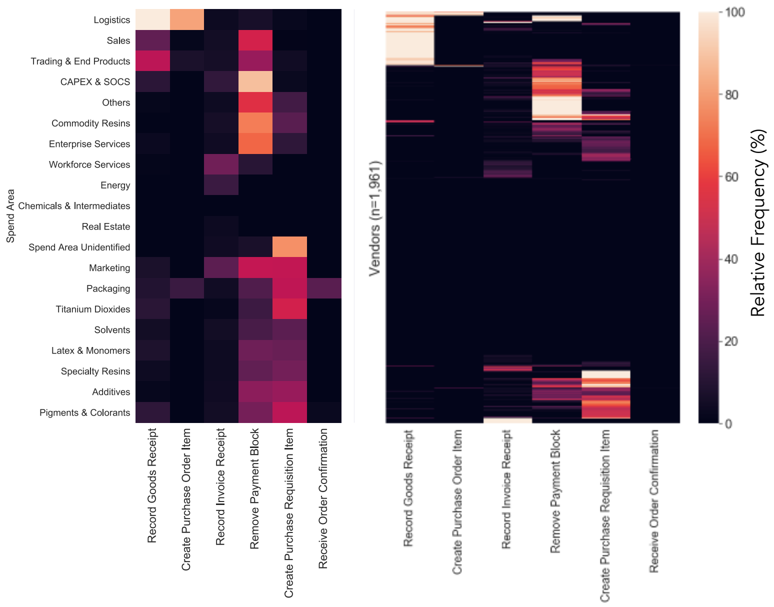

Fig. 6. Percentage of automation by spend area (left) and vendor (right)

5 An automatable event is defined as events within one of the 6 high-count activities shown in Fig. 5 executed by a 'user' resource (yellow). We assume any efforts to transform processes towards automation is best focused on frequently-occurring events.

Fig. 6 shows the distribution of such batch events by spend area and vendor. Brighter-colored tabs correspond to a higher percentage of events executed via batch processes by category. The following insights may be derived:

1. Automation is concentrated within certain spend areas and vendors. For instance, batch executions of 'Record Goods Receipt' are associated with vendors dealing in Logistics, Trading & End Products, and Sales spend areas. Processes in categories such as Real Estate and Chemical & Intermediates are relatively manual.

2. High automation rate for one activity is not necessarily correlated to a high automation rate for other activities across the same spend area or vendor. Cases in which goods receipts are consistently batch-recorded do not correspondingly see the same level of automation in purchase order creation, invoice receipt, or other events. Multiple separate efforts may need to be integrated in order to achieve a fully automated process.

4.3 Case Studies

We present the following vendor case studies to further illustrate how automation is utilized by the company.

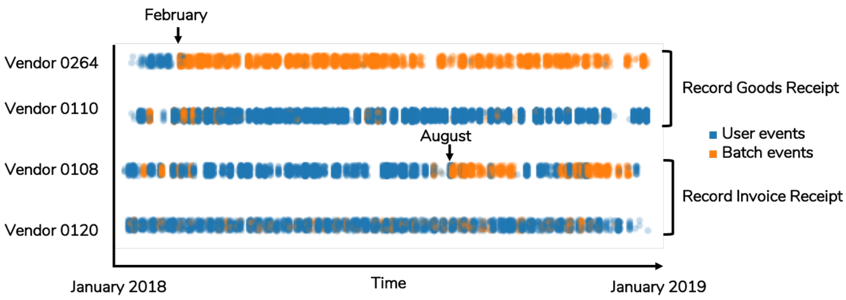

Fig. 7. 'Record Goods Receipt' and 'Record Invoice Receipt' events over time

Case Study 1: Vendor 0110 & 0264 (Record Goods Receipt). Of 3,802 instances of goods receipt records for items associated with vendor 0110, 13.7% were executed by batch processes, distributed randomly over the course of the event log (**Fig. 7**). The lack of discernible differences between cases with user- and batch-executed events suggests that automation of goods receipt record should be explored with this vendor in order to potentially reduce resource time required/payroll costs.

On the other hand, there appears to be a temporal pattern in user- and batch-executed for items associated with Vendor 0264, with a shift to batch goods receipt in February 2018. It appears that the process associated with this vendor underwent a change through the introduction of automation. We believe that further investigation into this vendor could reveal insights into how other vendor processes could have their goods receipts further automated as well.

Case Study 2: Vendor 0108 (Record Invoice Receipt). Similarly, invoices from Vendor 0108, which are exclusively of the Sales spend area, began being recorded largely by batch systems beginning in August 2018. Insights beyond the scope of this dataset would allow for further understanding of automation (e.g. paper-based invoices may need to be entered manually into the procurement system in the absence of a functional image-recognition algorithm).

Case Study 3: Vendor 0299 (Create Purchase Order Item). Some purchase order items from this vendor were created manually, while others were generated by batch processes. Approximately 4% of items were deleted by users soon after they were batch-created. This is may represent a notable inefficiency in the process, and therefore should be investigated.

Additionally, the presence of a batch user in a case typically reduces the number of human users involved by 0.8 with high statistical significance (p-value = 0). We surmise that there may be multiple obstacles preventing the automatability of an event. We recommend that the client first investigate processes associated with vendors that are already partially automated, and transition towards automation where viable in order to potentially reduce payroll costs.

5 Throughput Analysis: Backlogs, Payment Timing and Process Complexity

5.1 Key Findings and Recommendations

Backlog Analysis. "Record Invoice Receipt" is the activity most in need of increased automation, or more resources, in order to speed up the completion of a three-way or two-way match.

Payment Terms. We recommend that the company transform its payment process so that invoices can be paid at least weekly, thereby enabling it to pay invoices in a more optimized manner. This could lead to the following benefits:

1. Reduced costs by taking advantage of early-payment discounts. This represents EUR9.16 million in potential savings to the company.
2. Better overall cash management by optimizing payments up to limits provided in contract terms.
3. Reduced costs from penalties/interest charges for late payments.

Vendor Process Complexity. We tagged vendors using a custom two-dimensional complexity metric that identifies which vendors are most associated with common process inefficiencies. We recommend that processes associated with these vendors be examined further to understand how they might be standardized and streamlined.

5.2 Backlog Analysis

Backlog is a useful metric to determine where a process may be getting stuck. For an activity (e.g. "Record Invoice Receipt"), backlog is defined as the number of cases that are waiting for "Record Invoice Receipt" to happen at any given time. On the process map, these cases are the ones in the process of traversing arrows from other activities to "Record Invoice Receipt." Since the backlog for an activity varies over time, we examine the median and maximum values over the course of the event log time period.

Table 5. Backlog for first cycle of complete cases[6]

Backlog (cases)	3-Way After	3-Way Before, GR before IR	3-Way Before, IR before GR	2-Way	Consign.
GR Median	467	7,802	151	n.a.	876
GR Max.	762	9,802	396	n.a.	1,250
IR Median	949	9,249	739	8	n.a.
IR Max.	1,626	14,785	1,655	16	n.a.
CI Median	1,018	22,136	2,379	9	n.a.
CI Max.	1,640	30,533	4,609	56	n.a.
RPB Med.	n.a.	2,650	184	n.a.	n.a.
RPB Max.	n.a.	5,295	406	n.a.	n.a.

The generally high backlog at "Clear Invoice" – and how to address it – is discussed in depth in the payment-terms section of this report. In this section we highlight the relatively elevated backlogs at "Record Invoice Receipt," especially for the 3-way-after subprocess. Indeed, across the subprocesses that include this activity, median wait times between "Record Invoice Receipt" and the prior step range from 14 to 22 days. Both batch and human users perform this activity throughout the week, so one might wonder about the presence of high wait times and backlogs. One reason is that some cases are waiting for a "Vendor creates invoice" event to move on from "Create Purchase Order Item" or "Record Goods Receipt" to "Record Invoice Receipt." However, in the vast majority of cases "Vendor creates invoice" is already in place, at which point the process should be ready to proceed to "Record Invoice Receipt." We conclude that "Record Invoice Receipt" should be the focus for further automation. More specific opportunities for backlog/time reduction are as follows:

[6] Goods Receipt (GR), Invoice Receipt (IR), Clear Invoice (CI), Remove Payment Block (RPB)

Three-way matching, invoice after goods. Repetitions of "Record Invoice Receipt" and "Record Goods Receipt," which are sometimes extremely quick but can take a couple days, seem to be related to the Service segment (see discussion in compliance section).

Three-way matching, invoice before goods. Potential problem areas for investigation include the transition from "Record Goods Receipt" to "Remove Payment Block." Many unnecessary "Remove Payment Block" actions occur (see description of four subprocesses), but even those that are necessary have a median transition time of 9 hours, and the average transition time is 4 days, meaning that some of these cases get stuck on other activities for several days. This seems like a good candidate for further automation as there should be no obstacle to "Remove Payment Block" once goods have been received. A "Change Quantity" event can add a week to the median 8- to 10-day transition from "Create Purchase Order Item" to "Record Goods Receipt," so we recommend attempts to reduce the frequency of change events (see vendor complexity section).

Two-way matching. The largest median backlog by far (110 cases) sits at the "hidden" activity "Change Approval for Purchase Order." The main onward transition to "Record Invoice Receipt" takes only 3 days on average, so the problem seems to be transitions from "Change Approval for Purchase Order" back to itself, which happens more than once per case and takes 17 days on average, warranting an investigation into how many of these are necessary.

5.3 Payment Terms Analysis

Payment terms are a critical aspect of every purchase order. A purchase order constitutes a contract between the company and the vendor, which specifies the obligations that must be met by the respective parties. In general, once the vendor has received the purchase order from the company, it will deliver the goods or services and send out a corresponding invoice. The company must then send a payment within a period of time specified in the purchase order. Usually the payment deadline is 30, 60 or 90 days. If the payment is not accomplished in time, late fees may be charged to the company.

Cash management is an important consideration for any business, and consequently each side of the payment leg of the transaction has its own interest pertaining to the timeliness of its execution. From the company perspective, payment should be made as close to the deadline as possible to maximize its own cash position. Vendors would like to receive payment as soon as possible. In certain cases, suppliers will discount early payments as an incentive to vendors to pay quickly.

In this section we analyze the elapsed time between events constituting possible triggers of a "countdown clock" to the payment deadline, and the ultimate time at which the payment was made. These analyses consider only "completed" cases. We focus our analysis on the "3-way-before" process because it is the predominant process for transactions in this dataset.

Invoice Clearing Occurs at Set Intervals, While Vendor Invoicing Is Continuous. We consider "Vendor creates invoice" and "Clear Invoice" to be the typical start and end points of the payment timeline. **Fig. 8** shows the distribution of elapsed time between these two events by vendor. The heatmap shows that durations are clustered around certain bands for different vendors.

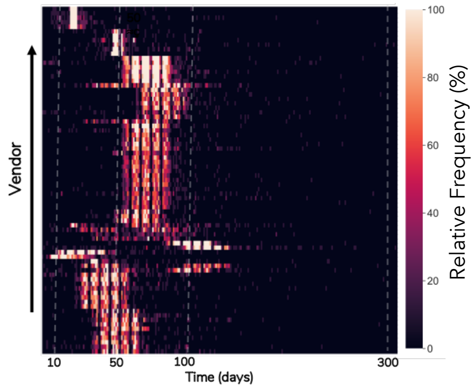

Fig. 8. Distribution of Elapsed time between "Vendor creates invoice" and "Clear Invoice" for 3-way-before, with vendors hierarchically clustered (vendors with at least 100 cases).

Looking at hierarchical cluster maps of "Vendor creates invoice" across time vs. "Clear Invoice" across time, we notice a pattern that explains the spread of durations for each vendor and shows opportunity to improve the process, as shown in **Fig. 9**.

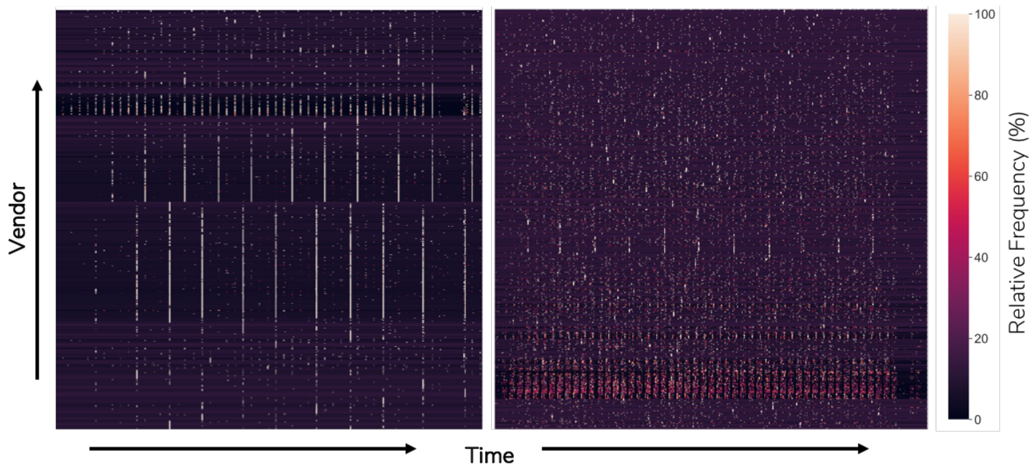

Fig. 9. "Clear Invoice" (left) and "Vendor creates invoice" (right) across time and vendor for 3-way-before, with vendors hierarchically clustered.

For most vendors, “Vendor creates invoice” events occur relatively constantly across weekdays, yet “Clear invoice” events are highly concentrated usually on the first and last Thursday of each month, as displayed by the light streaks. We observe that some vendors clear invoices on a weekly schedule, but overall, it appears that the “Clear Invoice” process is set by an inflexible and pre-determined schedule. This is suboptimal because it means that:

1. In the best case, vendor payments will be made too early, depriving the company of cash they may have been able to use or invest until the due date.
2. In the worst case, late payments if the “Clear Invoice” day is missed.

Below we inspect the distributions at the vendor level as well, which can give some information about typical payment terms for individual vendors.

Case Study: vendorID_0147. Vendor 0147 is chosen as an example case study for the type of analysis that can be performed. This particular vendor exhibits interesting characteristics:

1. Frequency: This vendor has the 37th most cases in the completed “3-way-before” process (952 completed cases).
2. Value: This vendor has a relatively high median PO Item value of EUR12,898 (82nd percentile).

Given this combination of frequency and high value, there is some opportunity for evaluating any potential costs associated with the company’s paying this supplier at any time other than the contracted due date. The data provides limited information about any due dates that would have been specified in the purchase order document, but by looking at the distribution of elapsed time between “Vendor creates invoice” and “Clear invoice,” we can gain some insights into the process, as shown in **Fig. 10**.

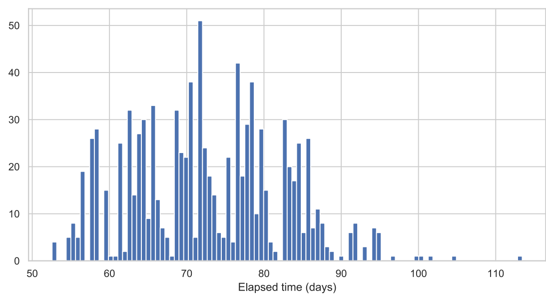

Fig. 10. Elapsed time between “Vendor creates invoice” and “Clear invoice” for completed cases associated with Vendor_0147

The figure shows peaks and valleys that fluctuate over time but do not appear to synchronize with expected payment terms of 30/60/90. Looking more closely at the occurrences of “Vendor creates invoice” and “Clear Invoice” for this vendor (**Fig. 11**), we observe that this action is performed once per month (with a few exceptions). Additionally, "Vendor creates invoice" occurs fairly evenly across weekdays. Consequently, we can conclude that the long amount of time that takes place between payment events can have a significant impact on the completion of a payment to a vendor. This is undesirable because it gives the company less flexibility in paying vendors on time, which could lead to penalties and interest charges. It also provides less ability for the company to use faster payment as leverage in negotiations over pricing or early-payment discounts.

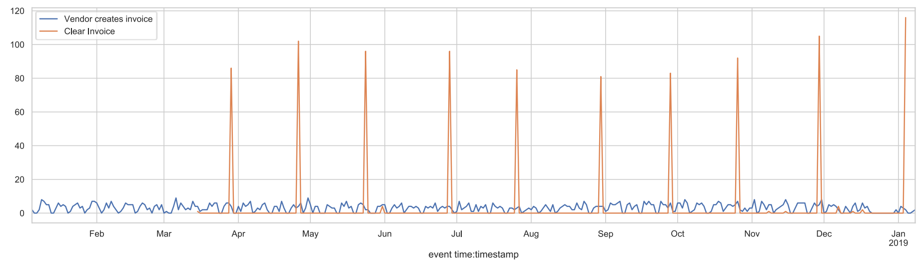

Fig. 11. "Vendor creates invoice" vs "Clear Invoice" across time for 3-way-before for Vendor 0147

Monetary Impact of Dynamic Invoice Clearance. Given insights from **Fig. 9**, revealing that the client currently pays most of its vendors on set dates (e.g. monthly or bi-monthly), we explored an alternative in which the client transitions into a more dynamic process during which an invoice is either cleared:

1. As late as is allowed by payment terms with its vendors, in order to optimize the company's cash flow.
2. Or, as early as reasonable in order to potentially negotiate an early-payment discount. Industry standard is often a 2% discount for paying within 10 days of the vendor's transmitting an invoice to the company [7].

We begin by quantifying the cumulative amount of bill-to-pay time historically taken up by simply waiting for invoice clearance. Bill-to-pay time is assumed to be the transition time between a "Vendor creates invoice" in a case log, to the first "Clear Invoice." Invoice clearance wait time is the duration between invoice clearance and any preceding major event, such as Record Goods Receipt.

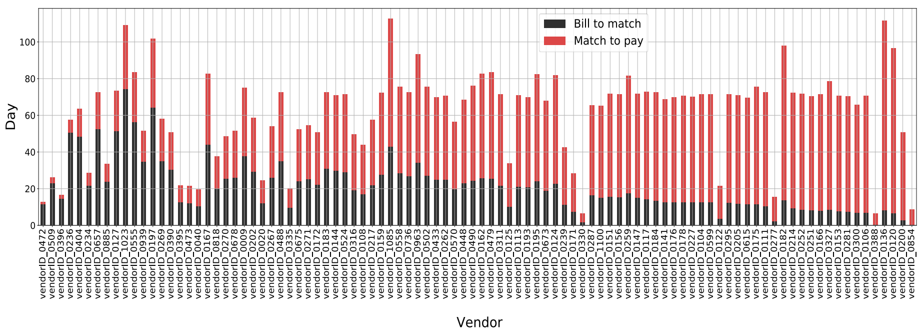

Fig. 12. Match-to-pay wait time (red) as a component of total bill-to-pay time for high-value vendors, sorted by fraction of red days.

Fig. 12 shows the median wait time to invoice clearance as a component of median total bill-to-pay time for high-value vendors with total item values worth over EUR1.5 million. We can see that the company pays certain vendors soon after a potential match is achieved across the purchase order, invoice, and/or goods receipt record. However, transactions with other vendors exhibit match-to-pay taking up the vast majority of process duration.

Since the data provided does not contain any information on payment terms, we cannot quantify the potential benefits of delaying payments to the limit specified by the terms. Instead, we attempt to quantify the theoretical total opportunity for savings if the company were to pay its invoices within 10 days to obtain a 2% discount on the value of the purchase item.

Within the scope of the available dataset, we assume that items that were paid more than 10 days after an invoice is created by the vendor, with goods received before or within 3 days of invoice receipt, would have been eligible for a 2% discount. Such completed cases correspond to 143,056 items, or 57% of items in the dataset. Taking the total value of those cases, a 2% discount represents EUR9.16 million in maximum potential savings to the company.

Note that this analysis could be altered to model more customized early payment schemes, various discount rates, or any opportunity costs to optimize the client's cash flow, if details of payment terms the client has negotiated with its vendors are provided.

We recommend that the client consider a transition towards a higher frequency invoice payment schedule (e.g. weekly, or even daily) where invoice clearance wait time is minimized when possible, or otherwise optimized with respect to the company's cashflow. An ability to pay more frequently may also give the company better overall negotiating power over pricing with vendors. While there are costs associated with such changes (e.g. restructuring the workflow of payroll departments or efforts associated with the computation of optimal invoice clearing), our preliminary calculations show great potential for cost savings from the ability of the company to negotiate and execute on early payment discounts.

5.4 Vendor Process Complexity

Quantifying the complexity of processes associated with certain vendors can lead to insights that may help the company identify opportunities to improve suboptimal processes. In this section we attempt to rank vendors based on their process complexity by looking at a combination of total vendor cases and number of process variants. This analysis looks at complexity in a holistic manner, by encompassing all events that happen in a given case for a given vendor.

Methodology. Once again, we focus on the "3-way-before" process. To evaluate complexity, we perform the following actions.

1. Process variants: We calculate the number of unique activity sequences completed for each vendor (ordered and unordered).
2. Total cases: We count the total number of cases associated with each vendor.
3. Complexity metric: We divide the number of cases by the number of unique variants for each vendor.
4. Subset high-opportunity vendors: We focus on vendors whose total case count is above the 90th percentile to maximize the chances that any process improvement for those vendors will have a meaningful overall impact.

Table 6. Top 5 Most Complex Vendors (90th percentile for Total Cases)

Vendor	Total Cases	Unique Variants	Cases per Variant
vendorID_0673	408	178	2
vendorID_0193	255	87	3
vendorID_0502	422	134	3
vendorID_0144	302	85	3
vendorID_0183	296	82	4

Table 6 shows that for the most complex vendors, there are only approximately 2-4 cases per variant. For comparison, the most consistent vendors in the 90th percentile have a case to variant ratio of 40 or more (with the most consistent vendor, vendorID_ 0550 displaying a ratio of 108) as shown in Fig. 18 in the appendix.

Analysis. Upon inspection of the processes associated with the problem vendors, we notice that many of these processes exhibit certain commonalities. Processes often include many "Change" type events, such as "Change Quantity" or "Change Price." While these activities are perhaps necessary, we believe that managers should investigate to ensure that needless complication is not occurring on the company side. We also note that these events tend to lengthen processes by contributing to backlog, as discussed in the prior section. These events may also be driven by the vendors. Further action could include working with vendors to help minimize these "noise" events, or even switching to vendors who may have more stable and efficient processes.

Fig. 13 explores the relationship between vendor complexity, case processing times, and vendor valuation. Each bubble represents a vendor at the top 90th percentile by case count. Time to resolve (TTR) is assumed to be the time between "Vendor creates invoice" and the activity preceding "Clear Invoice" to avoid the effect of any confounding variables in the analysis (e.g. certain invoices experienced a lot wait time to clear, certain items are created without work being done on it).

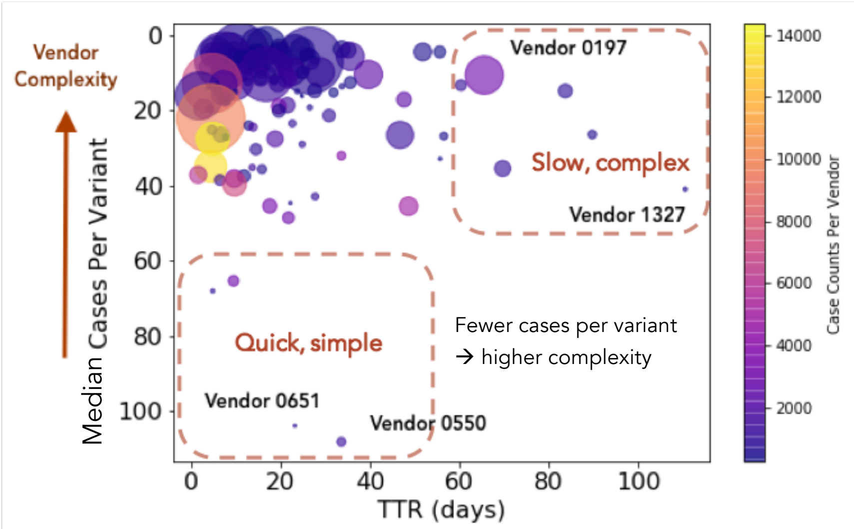

Fig. 13. Vendor complexity, median time to resolve (TTR), vendor valuation (cumulative amount of all purchase order items, represented by the size of the bubble), and case counts per vendor.

Vendor complexity moves inversely with the number of cases per variant (CPV). A higher CPV implies that a vendor has a standardized flow of activities to process a high number of cases and is thus less complex. The following insights can be derived from this analysis:

1. Many high-volume and/or high-value vendors such as vendor 0136 and vendor 0104 tend to be relatively complex. This may be thought of as a natural interpretation of complexity; vendors that are associated with a greater number of cases are likely to have a wider variety of unique process sequences, and items that are valuable may need to be processed with greater caution. We can note, however, that higher complexity does not necessarily translate to longer processing times, as seen towards the top left quadrant of the plot.

2. Cases from vendors such as 0197 may be problematic, given the relatively high complexity, time, case count, value, and resolution time. In relation to an earlier analysis on the monetary impact of dynamic invoice payments, high processing times may have an additional opportunity cost of lost potential to negotiate early payment discounts with vendors.

6 Social Network Analysis

6.1 Key Findings and Recommendations

User Relationship Analysis. A similar-task miner leads to the insight that, in 3,384 complete cases worth EUR61.4 million (1.7% of complete cases by number but 8.6% by value), the same human user performs "Create Purchase Order Item" as well as at least one instance of "Record Goods Receipt." This may raise the potential for fraud or errors by limiting checks and balances. Other combinations of compliance-related steps by the same human user are also present and may warrant further investigation. In order to direct the company's attention to other potential anomalies, we also highlight human users who seldom work together, hand off work or subcontract work.

User Activity Patterns. As expected, resources specialize in particular tasks and work Monday to Friday during business hours. The most prolific users who perform "Clear Invoice" seem to concentrate their activity heavily on Thursdays, which confirms a pattern observed in the payment terms section of this report that may need to be adjusted. We also flag how some human users work at unusual weekday hours or on Saturday mornings while others seem to schedule batch-like activity to occur overnight.

6.2 User Relationship Analysis

Segregation of Duties. Aside from the control-flow perspective on processes, valuable insights can be gained by looking at the resources (human users or batch systems) associated with each event. This is known as social-network analysis ("SNA"). SNA can be performed by using the resource column instead of the activity column in the event log when discovering a process map.

Since the data contains over 600 resources, the resulting diagram is too complicated to reasonably depict the flow of cases among resources. Therefore, we decided to use ProM for two different approaches, focusing on resources that performed more than 3,000 compliance-related events ("Create Purchase Order Item," "Record Goods Receipt," "Record Invoice Receipt," "Remove Payment Block" and "Clear Invoice") in complete cases (since these are the cases where payment was actually made). These 112 out of 627 resources account for 84% of all such events.

Segregation of duties within the purchasing process helps prevent errors and fraud as multiple resources check each other's work on a case. Using SNA, we identify human resources who are performing compliance-related activities that should be segregated, as well as several who operate more on-their-own than most, also a potential red flag. Finally, we note that no resource ("NONE") is recorded for 4% of all "Clear Invoice" events (7% by value). This percentage can be as high as 14% for some common sub-spend areas such as Packaging. "Clear Invoice" resources should always be recorded for better control of payments.

Similar-Task Miner. The first approach involves a group of ProM SNA modules (the second, involving dotted charts, is discussed in section 6.3 below). We began with the similar-task (ST) miner, which organizes groups of resources within "roles" (similar mixes of activities). We found that correlation (rather than Euclidean distance, similarity coefficient or Hamming distance) within the resource-activity matrix provided the best separation of roles.

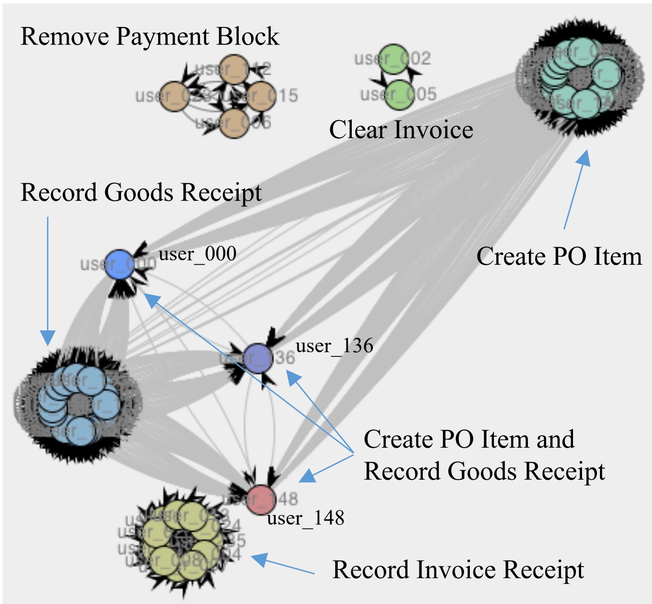

Fig. 14. Similar-task social network of human users performing over 3,000 compliance-related events (all complete cases)

In **Fig. 14**, the isolated circles for user_000, user_136 and user_148 represent users who most frequently perform both "Create Purchase Order Item" and "Record Goods Receipt" in the same case, alerting us to the presence of this potentially risky behavior. There are 44 human resources who do this though some are not active enough to appear in the diagram. As one would expect, the three problematic users are shown between a cluster of users who largely perform "Record Goods Receipt" and those who largely perform "Create Purchase Order Item." The other clusters largely perform "Clear Invoice," "Remove Payment Block" and "Record Invoice Receipt."

Breakdown of Problematic Cases. Of the 3,384 cases where the same human user performs "Create Purchase Order Item" as well as at least one instance of "Record Goods Receipt," 1,256 have Item Type Service, which is 49% of complete Service cases. This compares very unfavorably to the 0.8% of Item Type Standard cases that are problematic. Since all Service cases are 3-way-after, 3-way-after's percentage of problematic cases is also elevated at 14% (vs. 0.1% for 3-way-before).

The Spend Areas with the highest percentage of problematic cases are Workforce Services (79%), Enterprise Services (71%), Logistics (35%), Marketing (14%) and CAPEX & SOCS (9%). The most problematic vendors with more than 100 cases are vendorID_0003 (100%), vendorID_0000 (100%), vendorID_0741 (79%), vendorID_0277 (67%) and vendorID_1466 (19%).

Less-Connected Resources. Other ProM SNA modules that give insight include Hand-over of Work ("HoW"), Working Together ("WT") and Subcontracting ("SC"). These all indicate that resources in general hand off work (activities directly follow), work together (resources working on the same case) and subcontract (perform an activity both before and after another resource) with any and all other resources. Below we show some resources that tend to be less connected in these senses to other users so that the process owner can investigate whether reduced checks and balances with respect to these system users represents a risk (see appendix for graphical representations of these social networks):

- WT: user_005, user_087, user_136, user_171, user_186
- HoW: user_082, user_121, user_157, user_200
- SC: 40 of the 112 most active users do not subcontract at all

6.3 User Activity Patterns

The second approach uses dotted charts to compactly show when various resources perform various activities. This reveals the basic fact that resources tend to specialize in one particular activity while infrequently performing others. It also shows the weekly and daily work pattern where most activities by human users take place during business hours (around 8am-6pm weekdays). Unsurprisingly, batch resources work around the clock and on weekends.

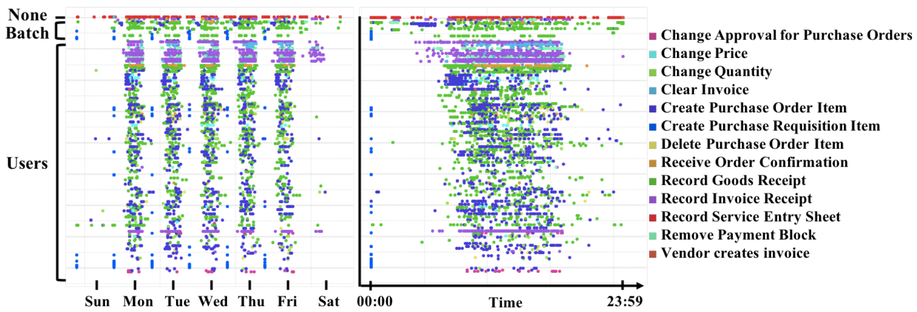

Fig. 15. Timing of most-frequent activities during the week and day by resource (those performing over 3,000 events)

Certain users seem to schedule "Create Purchase Requisition Item" to take place at midnight or 1am on Sunday-Thursday nights. This activity becomes much more common starting in September 2018, with several resources specializing in this activity added at that time. Some users concentrate their activities in the first few hours of the day, tapering off in the afternoon. A few users start work very early in the morning while others tend to arrive later but work into the evening. Finally, other users, often those performing "Record Invoice Receipt," tend to work on Saturday mornings. User_002 and user_005 are by far the most active "Clear Invoice" specialists, but user_002 does not work on Mondays, and user_005 does not work on Mondays or Wednesdays, perhaps contributing to the delays discussed in the payment terms section of this report.

7 Conclusion and Next Steps

In this report we explored a number of lines of inquiry which we believe have the potential to help the company enhance its processes, lower risks, and realize cost savings. We recommend that the process owner take steps to modify its payment processes to enable more dynamic payments to vendors. We also believe the process owner should investigate the noncompliant processes highlighted in the report to address potential risks. Further automation of processes could lead to a reduction in payroll costs. Monitoring vendors associated with high process complexity could lead to simpler and more efficient operations. Finally, our social network analysis points to instances where the segregation of duties is compromised and should be reviewed.

The data provided in the competition certainly gave us an intimate look into the way the subject company manages its purchase orders. However, this dataset only provides part of the picture. Organizations, especially complex ones, generate data in myriad ways. We believe that many of the findings in this report could be significantly enhanced by looking at other kinds of data produced by the subject company, such as employee rosters, inventory data, warehouse shipment delivery records, financial transactions, and even purchase order contracts with various vendors. Bringing disparate data sources together is a powerful way to gain a fuller understanding of what is actually happening within the enterprise at various levels. Additionally, we would also attempt to perform research on how the data is collected by speaking with users of the systems in question. Fundamentally understanding data collection is key to building a solid foundation for any analysis, and it provides solid ground for providing robust recommendations to decision-makers.

Appendix

Data Description.

Of the twenty-two columns in the dataset, we can immediately set aside five as not adding information:

- "case Purch. Doc. Category name": One value: "Purchase order."
- "case Source": One value, "sourceSystemID_0000."
- "case GR-Based Inv. Verif." and "case Goods Receipt": These True/False indicators, taken together, indicate which of the four main sub-processes a case follows, but "case Item Category" (see description below) already does this directly.
- "case Goods Receipt": See above.
- "event User": This column contains the same values as "event org:resource" (see below).

Of the remaining seventeen columns, five can vary within a case, i.e. they are properties of each event:

- "eventID": 1,587,802 values. Unique identifier for each event. We do not drop events that are duplicates aside from eventID.
- "event concept:name" (also known as "Activity"): 42 values. The action performed. Most common in descending order (10-20% of all events each): "Record Goods Receipt," "Create Purchase Order Item," "Record Invoice Receipt," "Vendor Creates Invoice," "Clear Invoice" and "Record Service Entry Sheet."
- "event org:resource" (also known as "Resource"): 627 values. The person or system who performs the "Activity." 25% of values are "NONE." Twenty non-human "batch" resources, the most common of which is "batch_06" (2.4% of events). 606 human users, the most common of which is "user_002" (10% of events).
- "event time:timestamp" (also known as "Complete Timestamp"): 166,419 values. The time/date when an event happens. Often several events within a case occur at the exact same time, which means that process maps discovered from this event log are accurate only up to a point as it is impossible to determine the "correct" order of simultaneous events.
- "event Cumulative net worth (EUR)" 25,164 values. The monetary value associated with each event. The distribution of these values by case – taken as the EUR value recorded upon a "Create Purchase Order Item" event – is highly skewed to low numbers (including zero for all of the Consignment cases) but ranges up to EUR8.8 million. The median purchase-order item is worth around EUR500. There are entries of amounts up to EUR28,994,530 associated with other activities in the log, but they do not match purchase-order item values and may be incorrect.

Finally, twelve of the columns are case variables, i.e. they are the same for every event in a given case:

- “case concept:name” (also known as “Case ID”): 251,463 values. The identification number for each case/purchase order item. It starts with the number contained in “case Purchasing Document,” followed by an underscore (“_”), followed by the number contained in “case Item.”
- “case Company”: 4 values. The subsidiaries of the coatings/paint company that have purchases in the log. “companyID_0000’ dominates with 99.6% of events. Almost all of the remaining 0.4% of events belong to “companyID_0003.” These 5,758 events (1,027 cases) are precisely those that have “case Item Category” as “2-way match” as well as those that have “case Item Type” as “Limit.” This group of cases has its own exclusive set of vendors while the other three sub-processes have overlap among their vendors. Finally, this group of cases all have “case Document Type” as “Framework order,” but there are many other “Framework order” events not in this group of cases. The few cases for “companyID_0001” (9 events) and “companyID_0002” (6 events) seem to be standard 3-way-match purchase orders.
- “case Document Type”: 3 values. Consistent across all cases/items within a “Purchasing Document.” “Standard PO” for 96.9% of events. “Framework order” accounts for 1.8% of events (see discussion of “companyID_0003” above). The remaining 1.4% of events are “EC Purchase order” and make up precisely the 1,425 cases that involve activities with “SRM” in the name. SRM stands for “Supplier Relationship Management,” an SAP system for efficient production of purchase orders in the field.
- “case Item”: 490 values. The item number within a Purchasing Document (see “Case ID” above). Item numbers tend to follow the convention 00010, 00020, 00030 … 00100, 00110 ... for each consecutive item in a Purchasing Document. In several documents item numbers deviate from this pattern (e.g. 53XXX7 from vendor 0197, 111X9 from vendor 0262, and 0000X from certain Service vendors). We encourage the company and vendors to adhere to a common numbering scheme for simplicity.
- “case Item Category”: 4 values. Indicates which of the four main sub-processes (2-way match et. al.) the case belongs to: “3-way match, invoice before GR” (78% of events), “3-way match, invoice after GR” (20%), “Consignment” (2%), and “2-way match” (0.4%). See compliance discussion for a description of how we break these sub-processes down further.
- “case Item Type”: 6 values. “Standard” for 78% of events. “Service” (16%) seems to be a special sub-process within “3-way match, invoice after GR” (see compliance section). Item Type “Consignment” labels exactly the same cases that have “case Item Category” as “Consignment.” “Limit” is discussed above under “case Company.”
- “case Name”: 1,886 values. This column pertains to vendors (who is selling the items to the company), but for vendors we decided to use the closely related but slightly more detailed column “case Vendor” (described below).
- “case Purchasing Document”: 76,273 values. The first part of the “case concept:name” identifier: unique 10-digit numbers, beginning with 2 (for Document

Type “EC Purchase order”) or 4 (Document Types “Standard PO” and “Framework order”). While 63% of documents contain only one item, a purchase document can have as many as 429 items, which tend to be created either at or around the same time. Within a purchasing document, items can be associated with multiple spend areas or item types but with only one vendor, document type or company. Service orders are largely single-item orders.

- “case Spend area text”: 21 values. The type of goods/services procured, top values being “Packaging,” “Sales” and “Logistics.” 1% blank entries.
- “case Spend classification text”: 4 values. A higher-level classification of goods/services procured. Same 1% of events blank as for the other “Spend” columns. “PR” (56% of events) stands for product-related, items that are raw material for the company’s products (accounted for as Cost of Goods Sold). “PR” comprises ten of the “case Spend area text” values (Packaging, Trading & End Products, Additives, Latex & Monomers, Solvents, Pigments & Colorants, Specialty Resins, Titanium Dioxides, Commodity Resins and Chemicals & Intermediates). “NPR” (42% of events) stands for non-product-related, e.g. capital expenditure or overhead, and comprises eight of the “case Spend area text” values (Sales, Logistics, CAPEX & SOCS, Marketing, Enterprise Services, Real Estate, Workforce Services and Energy). “OTHER” comprises two of the “case Spend area text” values (Others and Spend Area Unidentified).
- “case Sub spend area text”: 136 values. A lower-level classification of goods/services procured, most commonly “Products for Resale” (21% of events). Same 1% of events blank as for the other “Spend” columns.
- “case Vendor”: 1,961 values. Each vendor in the “case Vendor” column maps to exactly one vendor in the “case Name” column, but since there are 75 more values of “case Vendor,” each vendor in the “case Name” column maps to one or more vendor in the “case Vendor” column. The most extreme example of this is “case Name” vendor_0143, which maps to 11 different vendors in the “case Vendor” column.

Compliance Analysis: Random Forest Classifier.

The high dimensional nature of this dataset poses a challenge. Given so many categories of data to focus on, how do we narrow down which areas are worthy of analysis? A random forest classifier was used to help uncover patterns behind compliance. Given an input of event logs and a binary compliance flag, the classifier splits this input into random subsets and send each of them through a decision tree. Each tree then 'votes' on the outcome (i.e. compliant or non-compliant). These results are then aggregated, and the majority vote for each case of event is deemed the final prediction. The result of this classifier can be compared to the real compliance flag, and the model's key features (i.e. the most important characteristics of a case or event that delineates compliance) examined.

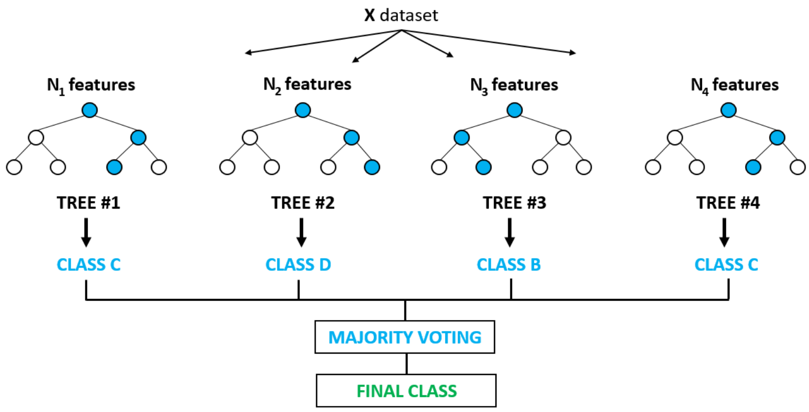

Fig. 16. Random Forest classifier. The dataset, a matrix containing event or case information, is split into *n* decision trees. Each tree votes on the best binary outcome. The votes are then aggregated, and the majority vote determines the final class outcome. [3]

Using a scikit-learn implementation of the random forest classifier [6], feature importance can be explicitly defined as a reduction in Gini impurity, or the likelihood that a datapoint is incorrectly labelled if it was classified according to the distribution of labels in that subset of the data, defined as [4].

$$G = \sum_{i=1}^{J} p_i(1 - p_i)$$

for a dataset with J classes, $i \in [1, 2, \ldots, J]$, and p_i being the fraction of items labeled with class i in the dataset. We want each split in the tree to be the most informative and delineating as possible. Feature importance (FI) is represented by the weighted information gain at each node, defined as [5],

$$FI = \frac{N_t}{N} \times (G_{parent} - \frac{N_{tR}}{N_t} * G_R - \frac{N_{tL}}{N_t} \times G_L)$$

where N is the total number of samples, N_t the number of samples at each node, N_{tR} is the number of samples at the right child, and N_{tL} at the left. As the name suggests, the higher feature importance, the more important and information a characteristic of the data is. We utilized the classifier as a data exploration tool to help uncover meaningful data segmentations given a dependent variable, such as compliance. Note that we chose a random-forest classifier over other models (such as logistic regression) due to its more intuitive nature and higher accuracy.

Figures.

Fig. 17 shows the distribution of median elapsed time between "Vendor creates invoice" and "Clear Invoice" grouped by vendor. Peaks are observed at around 22, 50, and 70 days.

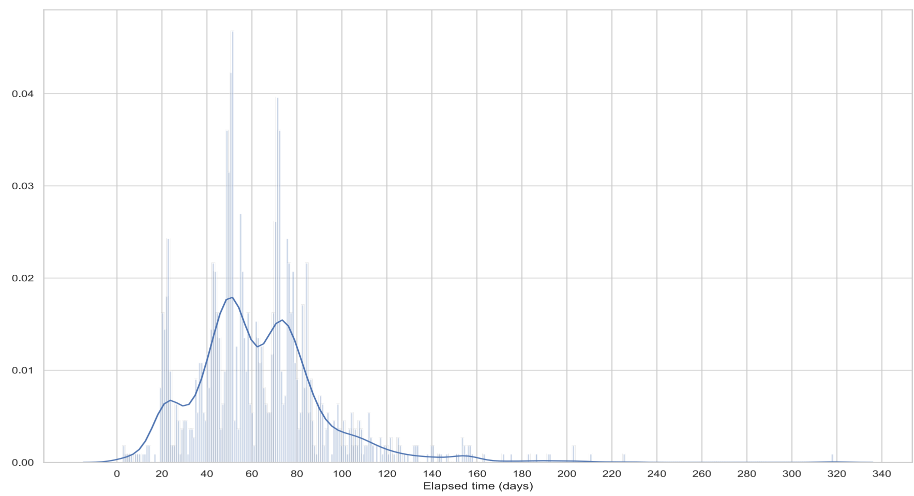

Fig. 17. Distribution of median elapsed time by vendor between "Vendor creates invoice" and "Clear Invoice" events.

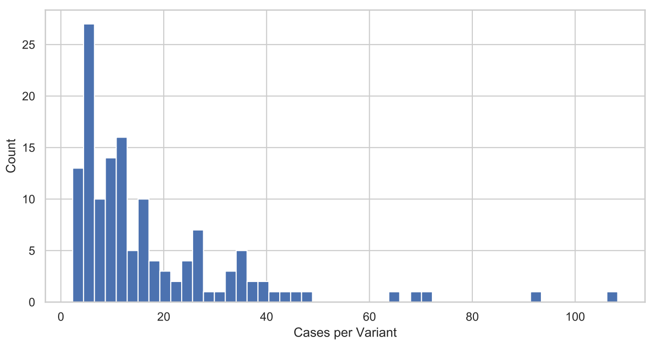

Fig. 18. Histogram of Cases per Variant by Vendor (90th Percentile of Total Case Count)

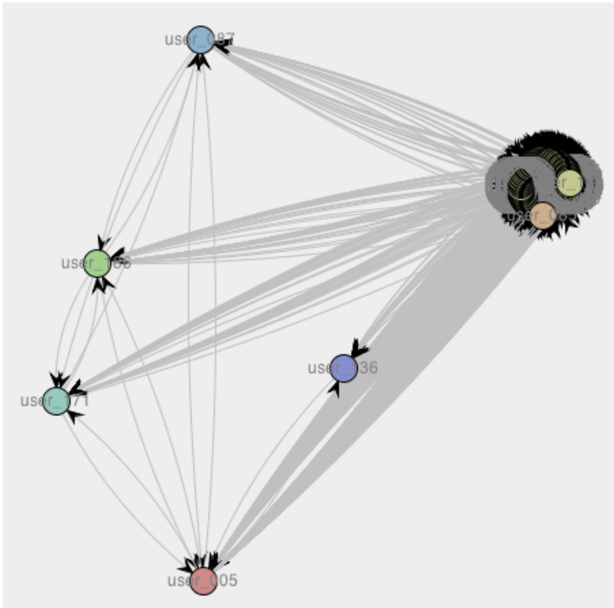

Fig. 19. Working-together social network of human users performing over 3,000 compliance-related events (all complete cases)

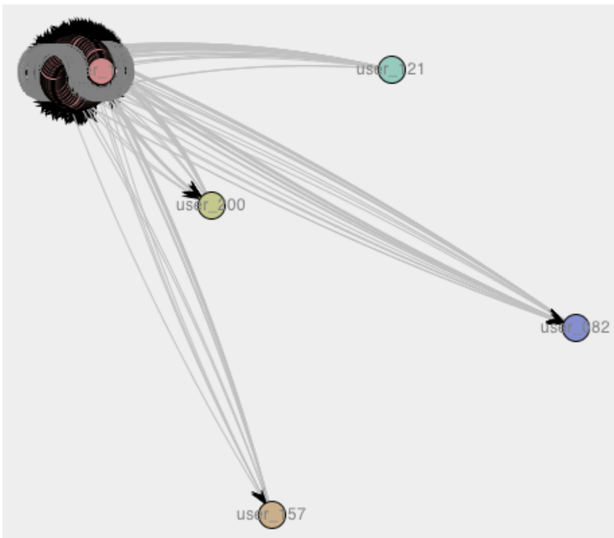

Fig. 20. Handover-of-work social network of human users performing over 3,000 compliance-related events (all complete cases)

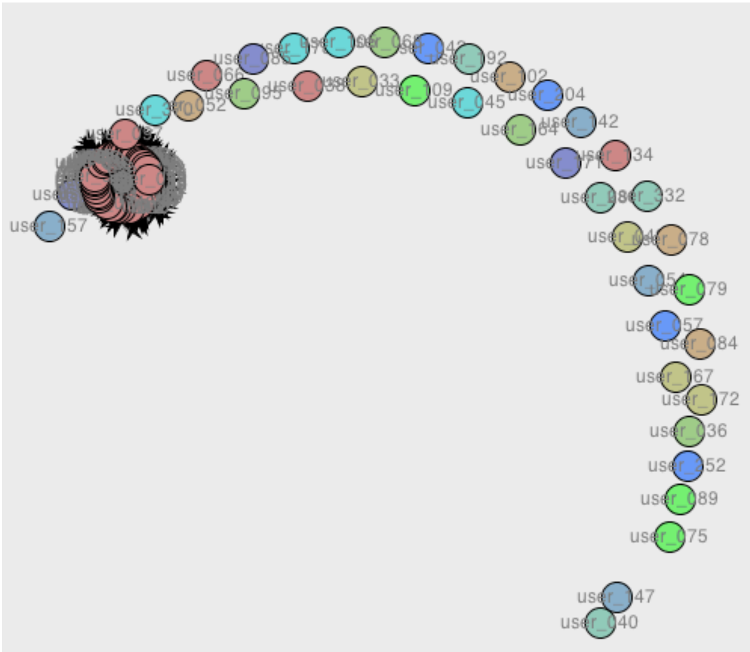

Fig. 21. Subcontracting social network of human users performing over 3,000 compliance-related events (all complete cases)

Tables.

Table 7. Case Completion Statistics (Expanded)

	Number	% by Number	Value (EUR millions)	% by Value	Mean Value	Median Value
Completed cases	196,881	73.2%	711.6	78.3%	3,615	491
Completed, with multiple invoice clearing[7]	6,774	2.7%	126.3	13.0%	18,638	1,251
Non-completed	54,582	26.8%	260.1	21.7%	4,766	565
Non-completed with Delete PO	8,561	3.4%	55.6	5.7%	6,500	1,125
Total	251,463	100%	971.8	100%	3,864	508

[7] "Clear Invoice" occurs up to 71 times in a case. These multiple-clear-invoice cases are part of all sub-processes except Consignment and are spread across all document types, item types, spend classifications etc.

References

1. van Dongen, B.F., Dataset BPI Challenge 2019. 4TU.Centre for Research Data. https://doi.org/10.4121/uuid:d06aff4b-79f0-45e6-8ec8-e19730c248f1
2. Process Mining Conference. BPI Challenge 2019. https://icpmconference.org/icpm-2019/contests-challenges/bpi-challenge-2019/
3. Global Software Support. Random Forest Classifier – Machine Learning. https://www.globalsoftwaresupport.com/random-forest-classifier-bagging-machine-learning/, last accessed May 7, 2019
4. Hauskrecht, M., Decision trees. https://people.cs.pitt.edu/~milos/courses/cs2750-Spring03/lectures/class19.pdf, last accessed May 7, 2019
5. Scikit-learn: Random Forest Classifier. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html, last accessed May 7, 2019
6. Pedregosa et al., Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.
7. Schaeffer MS. Essentials of accounts payable. John Wiley & Sons; October 15, 2002.