universiteit hasselt
KENNISCENTRUM
VOOR
ONDERNEMERSCHAP
EN INNOVATIE

Internal Fraud Risk Reduction by Data Mining and Process Mining: Framework and Case Study

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Toegepaste Economische Wetenschappen
aan de Universiteit Hasselt te verdedigen door

MIEKE JANS

Promotor
Prof. dr. N. Lybaert
Co-Promotor
Prof. dr. K. Vanhoof

Acknowledgements

Six years ago, I applied for the position of PhD-student. Thrilled to get started, but not really knowing what to expect, I came to the university on September 16, 2003. I knew I had the right mind set for doing research: I am critical (as I put in my resume, 'Critical to myself and others'), I always want to find an answer and never stop before I find an answer that satisfies me. Having the right mind set is a nice start though, but there are so many other factors (parameters if you want) that contribute to a successful realization of a PhD. A very important factor is the support of the people that surround you. I would like to grasp this opportunity to express my gratitude to a number of people for their time, advice, support, input and confidence.

First of all, I would like to thank my supervisor, co-supervisor and members of my commission: prof. dr. Nadine Lybaert, prof. dr. Koen Vanhoof, prof. dr. Roger Mercken en prof. dr. Eddy Vaassen for showing me how to conduct scientific research. Nadine, from the start you knew how to keep me enthusiastic about my research subject. You encouraged me to do research on items that triggered me, that took away my interest. And when I got carried away too much by some new inspiration, you fed my enthusiasm, but raised some critical objection at the same time to keep me down to earth. You were constructively critical when reading my texts and above all you believed in me for the two of us when I was not believing enough in myself. Thank you for being a supervisor with a warm heart. I am glad my research was further supported by my co-supervisor. Koen, your extensive knowledge on data mining and data analysis in general were a great help in the empirical parts of my study. I truly appreciate the comments you provided me of, the ideas to work further on and all the support you gave to me. I further like to thank prof. dr. Roger Mercken en prof. dr. Eddy Vaassen for their constructive comments and valuable suggestions. Towards the end of my PhD, two supplementary jury members joined my commission: prof. dr. Miklos Vasarhelyi of Rutgers University (New Jersey) and prof. dr. Jan

Vanthienen of K.U.Leuven, also providing me (again) of new insights and aspects I should (re)consider. I hope to have the opportunity to work with all of you in the future.

A special thanks goes to the people of 'Epsilon', my case company. My dissertation would have literally looked completely different without this cooperation. Mrs. F. Vandevoorde, I'd like to thank you for your believe in my project. From the first day on, you saw a clear win-win situation, which was our ultimate goal and which, I think, we successfully reached. I, as an academic, surely learned a lot from your business practice. I can only hope Epsilon learned as much from our joint project as I did. I would like to thank Marc to believe in me as strongly as his predecessor did. Not only did they encourage my work, they also made sure I had every help or collaboration within the organization I needed. Also Jérémy, who introduced me into the world of SAS and kept on helping me, even when not working any more at Epsilon: a big thank you. And Peter, I certainly want to thank you for your warm welcome, for guiding me everywhere and for helping me out with every issue that crossed my path. You made a real colleague of me, instead of some outsider that comes along once and a while. Thank you for that. And of course all other staff members of the Investigations department, too many to call by name, I want to thank for their help and their good fellowship.

Aside from my colleagues at my case company, I have plenty of colleagues at Hasselt University. I 'outlived' already a few generations of PhD students, every generation with its own characteristics. I have good memories on several activities, such as a trip to Scherpenheuvel, quizzes at Lieven and Lobke's place, coffee breaks with Bert and Stijn (Yannick kept on working), Efteling, Bobbejaanland, Blotevoetenpad and so much more. Not only were the extra activities a nice distraction from work, the work environment was great with such colleagues and help, if needed, was never hard to find. Thank you all.

Some academic colleagues, from outside Hasselt University, made a real difference to my work. At the annual congress of the European Accounting Association in 2007 in Lisbon, Miklos Vasarhelyi and Michael Alles had a first look at my preliminary work. They questioned everything I had done so far, let me rethink about my subject, and were as critical as one can be. At all following conferences, they kept on pushing me to a higher level. As is typical with critics: you never like them at the moment, but you are grateful for them afterwards. So a big thank you for this. I am looking forward to our future cooperation at Rutgers.

At last, but certainly not at least, I want to thank my friends and family. My friends who had to listen to lists of set-backs, which are inherent to doing research, comforted me when necessary. But of course they also shared my joy at moments of success, such as the acceptance of a paper for a conference or a best paper award. I also truly thank my parents and brothers and my family in law for their support. Although it was mostly like Chinese to them, they tried to understand what I was doing and even more so what I was trying to do. They were proud on me, even on moments they shouldn't have. But this believe kept me going. Kris, you were and are a husband that should I be envied for by a lot of women. You motivated me, but also showed me how to relax from time to time. You know me better than I know myself when it comes to work. On top of this, you made it possible I had time to work by doing a lot of work at home, for which I am ever thankful. Without you, your support, advice and confidence, I am not sure I would stand where I stand now. Thank you.

Contents

List of Abbreviations

ACFE	Association of Certified Fraud Examiners
COSO	Committee of Sponsoring Organizations of the Treadway Commission
ERP	Enterprise Resource Planning
FD	Financial Document (in SAP)
GR	Goods Receipt
ID	Identification
IFR^2	Internal Fraud Risk Reduction
IR	Invoice Receipt
IS	Information Systems
KDD	Knowledge Discovery in Databases
PG	Purchasing Group
PI-ID Process Instance ID	
PO	Purchasing Order
PO-invoices	Invoices for goods, ordered by a PO
nonPO-invoices	Invoices for goods without PO
SPO	Strategic Purchasing Organization
TPC	Tactical Purchasing Cell
WFMElt	Work Flow Model Element

SAP tables

BKPF	Header information of a financial document
BSEG	Item information of a financial document
CDPOS	Item information of a change
CDHDR	Header information of a change
EKBE	Purchasing Order History
EKKO	Header information of a PO
EKPO	Item information of a PO

List of Tables

List of Figures

Samenvatting

Onderzoeksopzet

Fraude binnen bedrijven is een 'hot topic'. Iedereen kent wel voorbeelden, zoals Enron, WorldCom, Lernout & Hauspie, Société Générale, etc. Dit zijn allemaal voorbeelden van fraude die zich binnen een bedrijf afspeelt en die gepleegd wordt door mensen die intern aan het getroffen bedrijf verbonden zijn. Dit type fraude, interne fraude genoemd, is het onderwerp van deze thesis. Verschillende onderzoeken brengen schokkende cijfers naar voren omtrent de kost van interne fraude. Onder deze onderzoeken vinden we de rapporten van de Association of Certified Fraud Examiners (ACFE) en van Pricewaterhouse&Coopers (PwC). Volgens het onderzoek van de ACFE (2008) verliest een bedrijf gemiddeld zeven procent van haar jaarlijkse inkomsten aan interne fraude. De studie van PwC (2007) belichtte dat 43 procent van de ondernemingen die deelnamen aan het onderzoek, slachtoffer was van fraude. Het fenomeen fraude maakt hierbij geen onderscheid tussen haar slachtoffers: zowel grote als kleine ondernemingen zijn gekozen doelwitten.

Het probleem van interne fraude is niet enkel 'hot' in het nieuws wanneer een nieuw schandaal de pers bereikt, het is ook in de bedrijfswereld een belangrijk aandachtspunt voor alle organisaties. De audit standaard SAS 99 en Sectie 404 van de Sarbanes-Oxley wetgeving weerspiegelen deze zorg. De aanpak die in de praktijk wordt voorgesteld is gebaseerd op het COSO model voor interne controle. COSO staat voor Committee of Sponsoring Organizations of the Treadway Commission. Deze commissie was in 1985 belast met het onderzoek naar oplossingen om financiële rapporteringsfraude tegen te gaan. Uiteindelijk heeft dit geleid tot een algemeen model om interne fraude bij bedrijven tegen te gaan, namelijk het COSO interne controle model, een model dat tegenwoordig in de bedrijfswereld nauwgezet gevolgd wordt.

Nadat de mogelijke oplossingen voor interne fraude binnen de bedrijfswereld, zoals

het COSO model zijn besproken, wordt er een literatuurstudie uitgevoerd in deze thesis. Deze academische literatuurstudie omtrent detectie en preventie van fraude heeft als focus de analysetechnieken die hiervoor worden gehanteerd. Uit de gevoerde literatuurstudie komt naar voor dat 1) vooral externe fraude op een kwantitatieve manier benaderd is, 2) er hiervoor veel data mining toepassingen zijn gebruikt, met name voorspellende data mining technieken en 3) dat er geen onderzoek is gevoerd naar transactiefraude, een veel voorkomende vorm van interne fraude.

De combinatie van de urgentie van het probleem van interne fraude en het gebrek aan data analyse onderzoek in deze context (in tegenstelling tot een grote stroom literatuur omtrent data analyse om externe fraude te detecteren) leiden tot de onderzoeksopzet van deze thesis: het creëren van een kadermodel voor interne fraude risicoreductie. Risicoreductie omvat hierbij zowel fraudedetectie als fraudepreventie.

Het IFR^2 Model

Om tot een model te komen voor interne fraude risicoreductie, worden in de thesis het pad van zowel de bedrijfswereld als van de academische wereld met elkaar gecombineerd. Het gecreëerde model wordt het IFR^2 Model genoemd. De kern van dit model houdt in om, als invulling van de vierde component van het COSO model (controle activiteiten), beschrijvende data mining technieken te hanteren om data van een bedrijfsproces te analyseren. Deze beschrijvende aanpak zal leiden tot het detecteren van herkenbare patronen, maar ook van eventuele afwijkingen. Het zijn deze laatste afwijkingen die onderzocht dienen te worden. Algemeen gesproken, zijn er vier soorten verklaringen voor uitschieters. Men kan te maken hebben met:

- een perfect natuurlijke transactie die volgens de regels is gebeurd, maar die door haar onfrequent karakter opvalt. Denken we bijvoorbeeld aan de aankoop van een computer mainframe tussen de aankopen van CD-ROMs.
- een fout die gemaakt is, zonder kwaad opzet,
- het omzeilen van procedures, of
- een fraudegeval.

De uitschieters dienen allen in één van deze vier categorieën geplaatst te worden. Dit dient te gebeuren door een samenwerking van de business experts en de onderzoekers. De eerste categorie van uitschieters zijn niet het doel van ons onderzoek, daar het onderzoeken van 'juiste' gevallen (gevallen volgens de regels) niet bijdraagt tot fraude risicoreductie en het de onderneming onnodig tijd en bijgevolg geld kost. Categorieën twee en drie zijn daarentegen voor een onderneming wel van waarde om te onderzoeken. Beide categorieën dienen gezien te worden in het kader van fraudepreventie. Zowel procedures die omzeild worden als fouten die onopgemerkt blijven, zijn situaties die niet altijd veel kwaad doen, maar die in een later stadium omgevormd kunnen worden tot fraudemogelijkheden. Dit heeft te maken met 'opportuniteit', één van de drie elementen van de fraude driehoek. Indien opportuniteit al aanwezig is, en later worden de overige twee persoonlijke elementen (rationalisering en druk) geactiveerd, is de kans groot dat deze persoon gaat frauderen. De vierde categorie van uitschieters, fraudegevallen, is uiteraard van belang voor fraudedetectie.

In de thesis is later het IFR^2 Model uitgebreid met een ander onderzoeksdomein, namelijk dat van process mining. Dit leidde tot het Extended IFR^2 Model. Process mining heeft tot doel om, op basis van een event log van een bedrijfsproces, het werkelijk gevolgde proces te achterhalen. Dit proces kan namelijk afwijken van wat in de designfase van het bedrijfsproces werd vooropgesteld en is daarom interessant in het kader van fraudepreventie. Verder biedt process mining een nieuwe mogelijkheid op het gebied van het bewaken van interne controles, de vijfde component van het COSO model. Met process mining zijn we immers in staat om geval per geval te checken of aan een bepaalde interne controle voldaan is of niet. Zo kunnen we bijvoorbeeld voor iedere transactie die doorheen het systeem gegaan is, testen of deze wel degelijk aan de voorwaarde voldeed dat de eerste en de tweede goedkeuring door verschillende personen is gegeven. Deze test kunnen we bovendien voor elke transactie uitvoeren, wat een groot contrast vormt ten opzichte van de huidige manier van werken met random samples.

Gevalstudie

Het oorspronkelijke en het Extended IFR^2 Model zijn toegepast in een onderneming. De onderneming, die anoniem wenst te blijven, is actief in de financiële dienstverlening. In de thesis zijn allereerst de belangrijkste bedrijfsprocessen besproken, vervolgens is het aankoopproces geselecteerd en geanalyseerd om er het IFR^2 Model op toe te passen. Vooraleer met de data mining toepassing te starten is ook een risk assessment uitgevoerd om te kijken waar de grootste risico's zich in dit proces situeren.

De toepassing van het IFR^2 Model is in twee delen gebeurd. Een groep van recente bestelbonnen die door het aankoopproces zijn gegaan, en een kleiner groepje van oude bestelbonnen, zijn apart onderworpen aan een latent class clustering techniek. Bij beide groepen werden telkens drie clusters gevormd, waarvan er één heel klein was met slechts één á twee procent van de onderzochte bestelbonnen. Voor de groep van recente bestelbonnen was dit klein percentage echter nog steeds een te grote groep om handmatig te onderzoeken. In verder onderzoek zou hier een double loop expert systeem op geplaatst kunnen worden. Voor de kleinere groep met oude bestelbonnen was het wel haalbaar om de kleine cluster handmatig te onderzoeken. De cluster bevatte tien bestelbonnen, waarvan in negen gevallen de procedures waren omzeild en de overige bestelbon een fout bevatte. In het kader van fraude risicoreductie zijn dit twee gewenste categorieën van uitschieters om te vinden.

Ook het process mining deel van het Extended IFR^2 Model is toegepast op de gevalstudie in deze thesis. De resultaten hiervan tonen aan dat het process mining domein een zeer nuttige bijdrage kan leveren tot inzichten in processen en tot het al dan niet volgen van procedures. Ook de haalbaarheid om zulke tests op maandelijkse basis te verrichten werd positief geëvalueerd.

Bijdragen

Met deze thesis worden verschillende bijdragen geleverd. De belangrijkste bijdrage situeert zich op het vlak van academisch onderzoek. Voor de eerste keer wordt er op een kwantitatieve manier data geanalyseerd met als doel om interne fraude risico te verminderen. In tegenstelling tot het onderzoek naar externe fraude is de focus in deze studie zowel fraudedetectie als fraudepreventie, en worden er beschrijvende data mining technieken voorgesteld, in plaats van voorspellende technieken. Verder worden er met deze thesis ook bijdragen voor de meewerkende organisatie geleverd, voor de bedrijfswereld, voor software verkopers en voor de economie in het algemeen.

Chapter 1

Introduction

Saying that fraud is an important (however not loved) part of business, is nothing new. Everybody can recall some kind of fraud that has been all over the news. If it were Enron, WorldCom, Lernout & Hauspie, Ahold, Société Générale or another case does not matter. Fact is that fraud has become a serious part of our life and hence a serious cost to our economy.

There are several types of corporate fraud, the type of fraud which is the subject of this chapter. The most prominent distinction one can make in fraud classification is internal versus external fraud, a classification based on the relationship the perpetrator has to the victim company. Management fraud is an example of internal fraud, where insurance fraud is a classic example of external fraud. This dissertation will be oriented toward internal fraud, given the serious cost it represents and the lower representability of academic research in this type of fraud.

We start this chapter with a general section about corporate fraud, handling both external and internal fraud. This includes definitions and classifications of fraud, some important fraud theories, fraud in an agency context, the cost of fraud and some ideas about fraud detection and fraud prevention. In a following section existing literature concerning fraud detection and prevention is reviewed. This emphasizes the lack of research in internal fraud as opposed to external fraud. Also, this literature review brings the current methodology for external fraud detection to light. Especially the use of data mining techniques, with high emphasis on fraud detection, is very characteristic. In Section 1.3 the research objective is expounded.

1.1 Fraud

1.1.1 What is Fraud?

Fraud is deception. Whatever industry the fraud is situated in or whatever kind of fraud you visualize, deception is always the core of fraud. There are many definitions of fraud, depending on the point of view considering. For the most general meaning, *The American Heritage Dictionary, Third Edition* defines fraud as *'a deception deliberately practiced in order to secure unfair or unlawful gain'*. To fully articulate a case of fraud, Davia et al. (2000) paraphrase this in a number of items that must be identified:

- a victim,
- details of the deceptive act thought to be fraudulent,
- the victim's loss,
- a perpetrator (i.e., a suspect),
- evidence that the perpetrator acted with intent,
- evidence that the perpetrator profited by the act(s).

In a nutshell, *"Fraud always involves one or more persons who, with intent, act secretly to deprive another of something of value, for their own enrichment"* (Davia et al., 2000). Also Wells (2005) stresses deception as the linchpin to fraud. The kind of fraud as subject matter of his book is *occupational fraud and abuse*. This is a delineation of fraud, which is also periodically investigated by the Association of Certified Fraud Examiners (ACFE). In their *2008 Report to the Nation on Occupational Fraud and Abuse*, the ACFE defines occupational fraud and abuse as: "*The use of one's occupation for personal enrichment through the deliberate misuse or misapplication of the employing organization's resources or assets.*" (ACFE, 2008). This definition encompasses a wide variety of conduct by executives, employees, managers, and principals of organizations. Violations can range from asset misappropriation, fraudulent statements and corruption over pilferage and petty theft, false overtime, using company property for personal benefit to payroll and sick time abuses (Wells, 2005). Although this type of fraud encompasses many kinds of irregularities, mind that it does not cover all kind of frauds. Only internal corporate fraud is included. For example fraud against the government (non corporate fraud) or fraud perpetrated by customers (external fraud) are not included.

1.1.2 Classifying Fraud

The delineation of fraud to 'occupational fraud and abuse' is one way to categorize fraud. There are numerous other ways of classifying fraud. A classification that resembles however this first delineation, is the distinction Bologna and Lindquist (1995) make between *internal* versus *external* fraud. This classification, applied in the field of corporate fraud (fraud in an organizational setting), is based on whether the perpetrator is internal or external to the victim company. Frauds committed by vendors, suppliers or contractors are examples of external fraud, while an employee stealing from the company or a manager cooking the books are examples of internal fraud. What is seen as internal fraud, following this definition, is in fact occupational fraud and abuse, since one has to be internal to a company and abuse its occupation to commit internal fraud. We put internal fraud and occupational fraud and abuse as equivalents. A combination of internal and external fraud can also occur, for example when an employee collaborates with a supplier to deprive the company.

Bologna and Lindquist (1995) mention, in addition to other classifications, another way of classifying fraud: *transaction* versus *statement* fraud. The authors define statement fraud as *"the intentional misstatement of certain financial values to enhance the appearance of profitability and deceive shareholders or creditors."* Transaction fraud is intended to embezzle or steal organizational assets. Davia et al. (2000) distinguish two related types of fraud: *financial statement balance* fraud and *asset-theft* fraud. The authors state that the main difference between the former and the latter is that there is no theft of assets involved in financial statement balance fraud. Well known examples of this type of fraud are Enron and Worldcom. We see this classification (financial statement balance fraud vs. asset-theft fraud) as an equivalent of Bologna and Lindquist (1995)'s statement and transaction fraud. The ACFE categorizes internal fraud into asset misappropriation, corruption and financial statement fraud (ACFE, 2008). We put both asset misappropriation and corruption in the category of Bologna and Lindquist (1995)'s transaction fraud, because both these types of fraud are generally realized by forging some transaction.

Bologna and Lindquist (1995) give two more classifications of fraud - all classifying corporate fraud. A first classification is fraud *for* versus *against* the company. The former contains frauds intended to benefit the organizational entity, while the latter encompasses frauds that intend to harm the entity. Examples of fraud for the company are price fixing, corporate tax evasion and violations of environmental laws.

While these frauds are in the benefit of the company at first, in the end the personal enrichment stemming from these frauds are the real incentives. Frauds against the company are only intended to benefit the perpetrator, like embezzlement or theft of corporate assets. The authors draw attention to the fact that not all frauds fit conveniently into this schema, for example arson for profit, planned bankruptcy and fraudulent insurance claims.

A last distinction Bologna and Lindquist (1995) refer to is *management* versus *non-management* fraud, also a classification based on the perpetrator's characteristics.

These different classifications all present another dimension and can display some overlap. In Figure 1.1 we present an overview of how we see the different classifications and their relations to each other, hereby making some assumptions.

The most prominent classification is the internal versus external fraud, since all other classifications are situated within internal fraud. As already pointed out, we see occupational fraud and abuse as an equivalent of internal fraud.

Within internal fraud, three different classifications occur. We start with a distinction between statement fraud and transaction fraud, respectively financial statement balance fraud and asset-theft fraud in terms of Davia et al. (2000). A second distinction is based upon the occupation level of the fraudulent employee: management versus non-management fraud. We assume that managers can commit both statement and transaction fraud, yet non-management is in our view restricted to transaction fraud only. The last classification we introduce in this overview is fraud for versus fraud against the company. Although fraud for the company does not necessarily need to be statement fraud (for example breaking environmental laws), an overlap is realistic. With the classification for versus against, we again make an assumption. Contrary to fraud against the company, we believe only managers are in an advantageous position to commit fraud for the company, hence the overlap with only management fraud. Whereas fraud against the company is believed to be committed both by managers and non-managers. A last assumption is made concerning the nature of statement fraud. We assume all statement fraud is committed to improve the company's appearance and never to harm the company. Therefor we assume statement fraud is always profiled as fraud for the company, never against the company.

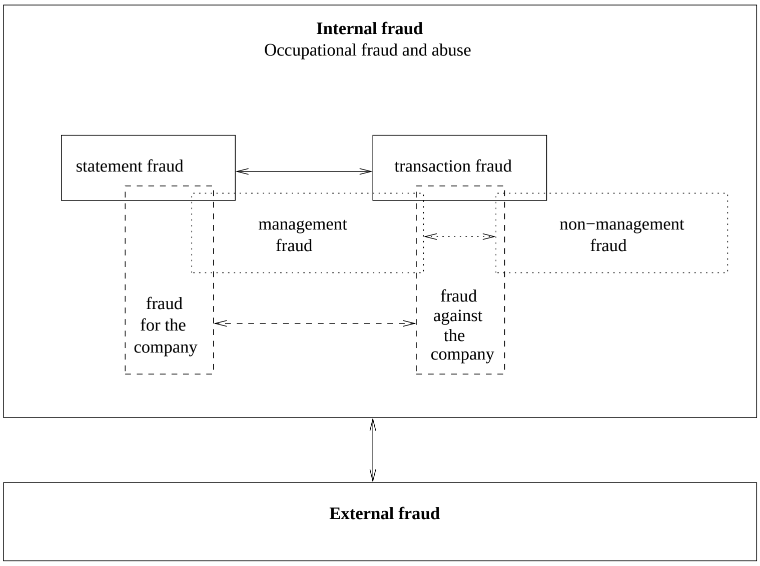


Figure 1.1: Fraud classification overview

1.1.3 Important Fraud Theories

Fraud was first most investigated from a social point of view. Edwin H. Sutherland was according to Laub (2006) undoubtedly the most influential criminologist of the twentieth century and it was Sutherland that coined the term *white-collar crime* in 1939 (Wells, 2005). Sutherland's most important contribution to the criminal literature however was his "Theory of differential association". (Sutherland et al., 1992) This theory has been a paradigm for the field of criminology (Laub, 2006). The differential association theory's basic tenet is that crime is learned. Sutherland theorized that people, as a result of more intimate, longer, more frequent, and more intense associations with cultural "definitions" favorable to criminal behavior as opposed to those unfavorable, learn criminal behavior. They learn the techniques to commit the crime, but also the attitudes, drives, and rationalizations (Sutherland et al., 1992; Tittle et al., 1986; Wells, 2005).

One of Sutherland's students during the 1940's was Donald R. Cressey. Cressey's dissertation concentrated on embezzlers and he interviewed about 200 incarcerated inmates at prisons in the Midwest (Wells, 2005). Cressey's hypothesis, better known as the "fraud triangle", sees three elements necessary for someone to commit fraud.

There has to be *pressure* (or a "perceived non-shareable financial need"), a perceived *opportunity* and the perpetrator must be able to *rationalize* its acts (Wells, 2005). The fraud triangle is cited many times in fraud literature and has become an important hypothesis.

Steve Albrecht, amongst others, adapted Cressey's fraud triangle from criminology to accounting. Albrecht was educated as an accountant, unlike Cressey and Sutherland (Wells, 2005), and was especially interested in factors that led to occupational fraud and abuse. Albrecht and two of his colleagues conducted an analysis of 212 frauds in the early 1980s, leading to their book, entitled *Deterring Fraud: The Internal Auditor's Perspective.* Albrecht et al. (1984)'s findings were similar to Cressey's. Also three elements were theorized to be present in the case someone commits fraud: *situational pressures*, *perceived opportunities*, and *personal integrity*, in fact the same elements as in the fraud triangle. They illustrated their concept by the "fraud scale". If the first two elements are high and the third low, the fraud risk is expected to be high (Albrecht et al., 1984). Choo and Tan (2007) even link Albrecht et al. (1984)'s findings with the *Statement on Auditing Standards No. 99: Considerations of Fraud in a Financial Statement Audit.*

In 2004, the fraud triangle theory was again adapted/elaborated by Albrecht (Albrecht et al., 2004). The fraud triangle theory was combined with the agency theory from economic literature and the stewardship theory from psychology literature and describes the 'Broken Trust' theory (a name that is put forward by Choo and Tan (2007)). The broken trust theory applies to a specific type of corporate fraud, namely corporate executive fraud. This theory explains corporate executive fraud in a matrix that links the three variables to corporate executives behavior, their compensation and the corporate structure.

Choo and Tan (2007) complement the broken trust theory with the 'American Dream' theory from sociology literature and support it with anecdotal evidence from three high profile executive frauds (Enron, WorldCom, and Cendant). Again, this theory only relates to corporate executive fraud.

Wolfe and Hermanson (2004) added a fourth element to the fraud triangle, resulting in the fraud diamond. This fourth element comprises someone's *capability* to turn an opportunity effectively into fraud. Without this capability, the fraud triangle elements will not result in fraud. Capability addresses the issue of the 'right' mind set to commit fraud. Besides the aspect of rationalization, the (future) perpetrator needs to be convinced he can 'beat the system'.

Two last persons we would like to mention because of their relevant research in this area, are Richard C. Hollinger and John P. Park. Hollinger and Park conducted in the early 1980's a survey of nearly 10,000 American workers. They reached a different conclusion than Cressey (and hereby also different than Albrecht et al.). They found that employees steal primarily as a result of workplace conditions. (Wells, 2005) Consequently, personal characteristics such as pressure and rationalization are irrelevant in accordance with this research.

1.1.4 Fraud as an Agency Problem

In the previous section, important and direct research about fraud is recapitulated. There is yet another stream of research, not directly brought in connection to fraud, which is nevertheless important enough to mention in this dissertation: the agency theory. The agency problem was introduced by Jensen and Meckling in 1976 and has triggered lots of research since then. An agency relationship was seen as *"a contract under which one or more persons (the principal(s)) engage another person (the agent) to perform some service on their behalf which involves delegating some decision making authority to the agent."* (Jensen and Meckling, 1976) The problem associated with this relationship stems from the believed divergence of the self-interested parties. The agents will tend to act in their own interest instead of the principal's interest. The fact of imperfect information between the two parties, called an information asymmetry between the principal and the agent, is the soil of this agency problem. This information asymmetry leads to two problems: ex ante there is the problem of *adverse selection*, ex post, the problem of *moral hazard*. The adverse selection problem refers to the disadvantageous situation the principal is in when selecting an appropriate agent. Moral hazard refers to the agents' advantage of his behavior, unobservable by the principal. The agent may not exert the promised effort or engage in hidden actions at the principal's expense. (Pavlou et al., 2007) For overcoming these problems, steps will be taken in the form of monitoring the agent (by the principal) and bonding (by the agent). The resulting expenditures are agency costs, as is the cost of incentives and punishments. Fama and Jensen (1983) define agency costs as *"the costs of structuring, monitoring, and bonding a set of contracts among agents with conflicting interests. They also include the value of output lost because the costs of* *full enforcement of contracts exceed the benefits."*

The principal-agent (P/A) model elaborated by Jensen and Meckling was primarily focussed on the relationship between the board of directors of large for-profit organizations and its managers, along with a self-interested behavior of its agents. The P/A perspective is however ubiquitous and can be applied to many types of relationships. In this light we see the problem of internal fraud as an agency problem, more specifically a problem of moral hazard. There are two possible P/A relationships in the context of internal fraud. The first relationship is the one between the board of directors as a principal and the managers as agents, the second is the relation between the managers as principals and the employees as agents. Moral hazard in the first relationship will more likely lead to statement fraud, while it in the second relationship more likely may lead to statement fraud. The first agency problem is mitigated by numerous (accounting) rules, the second problem by internal audits and internal control systems. The costs these controls bring forward are agency costs, stemming from this particular agency problem of internal fraud.

We do not intend to elaborate a model that estimates the agency costs of the internal fraud agency problem. It is however a very interesting framework to embed fraud research in. Boyer (2007) for example already put insurance fraud as an agency problem with two types of agents: the Truths and the Dares. The Truths always report the true state of the world, while the Dares dare to misreport the true state of the world. Boyer (2007) uses this framework to make some statements about allocating resources (or not) to fraud prevention.

1.1.5 Cost of Fraud: Some Numbers

Fraud is a million dollar business, as several research studies on this phenomenon report shocking numbers. Concerning internal fraud, two elaborate surveys, one conducted in the United States by the Association of Certified Fraud Examiners (ACFE)[1]

[1] *"The Association of Certified Fraud Examiners (ACFE) is the world's premier provider of anti-fraud training and education. Together with nearly 40,000 members, the ACFE is reducing business fraud worldwide and inspiring public confidence in the integrity and objectivity within the profession."* (www.acfe.com)

[2] (ACFE, 2008) and one worldwide by PricewaterhouseCoopers [3] (PwC, 2007), yield the following information about corporate fraud:

Forty-three percent of companies surveyed worldwide (PwC survey) has fallen victim to economic crime in the years 2006 and 2007. No industry seems to be safe and bigger companies seem to be more vulnerable to fraud than smaller ones. Small businesses however suffer disproportionate fraud losses. The average financial damage to companies subjected to the PwC survey, was US$ 2.42 million per company over the past two years. Participants of the ACFE study estimate a loss of 7% of a company's annual revenues to fraud. Applied to the 2008 United States Gross Domestic Product of US$ 14,196 billion, this would translate to approximately US$ 994 billion in fraud losses for the United States only.

Regarding to the types of fraud, asset misappropriation was number one in both studies. In the PwC survey, this was followed by financial misrepresentation and corruption, false pretences, insider trading, counterfeiting and money laundering. The ACFE report handles a different classification, where asset misappropriation takes 88.7% of the reported cases for its account, corruption 27.4% and fraudulent statements 10.3%.[4]

The numbers mentioned above all concern forms of internal fraud. There are however also large costs from external fraud. Four important domains afflicted by fraud are regularly discussed: telecommunications, automobile insurance, health care and credit cards. On these domains, we found the following numbers:

Globally, telecommunications fraud is estimated at about US$ 55 billion. (Abidogum, 2005) For the second domain, the automobile insurance fraud problem, Brockett et al. (1998) cite an estimation of the National Insurance Crime Bureau (NICB) that the annual cost in the United States is US$ 20 billion. At the web site of the NICB we read: "Insurance industry studies indicate 10 percent or more of property/casualty insurance claims are fraudulent." (NICB, 2008) Concerning health care insurance claims fraud, the United States National Health Care Anti-Fraud Association (NHCAA)

[2] 959 cases of occupational fraud, reported by a Certified Fraud Examiner late 2007, early 2008, are subject of this report.

[3] 5.428 companies (all PwC clients)across 40 countries around the world are subjected to the Global Economic Crime Survey 2007, a biennial survey conducted by PwC.

[4] The sum of the percentages exceeds 100% because several cases involved schemes that fell into more than one category

estimates *conservatively* that of the nations annual health care outlay, at least 3% is lost to outright fraud. This is $68 billion. Other estimates by government and law enforcement agencies place the loss as high as 10% of their annual expenditure. (NHCAA, 2008) Concerning the fourth domain, credit card fraud, Bolton and Hand (2002) cite estimates of US$ 10 billion losses worldwide for Visa/Mastercard only.

1.1.6 Prevention versus Detection

A lot is written about how to detect fraud. However many authors, like Bologna and Lindquist (1995), state that prevention should take precedence over detection. The authors mean by fraud prevention creating a work environment that values honesty. This includes hiring honest people, paying them competitively, treating them fairly, and providing a safe and secure workplace.

In the *Accountant's Guide to Fraud Detection and Control*, Davia et al. (2000) state that it is management's responsibility to allocate resources and emphasis to fraud-specific internal controls and to proactive fraud-specific examinations. These approaches are examples of fraud prevention. However, the authors point out that it is a mistake to think in terms of prevention versus detection, companies should invest in both. Strong internal controls as fraud prevention are very important, but they are best reinforced by following fraud-specific examinations.

In the above mentioned studies of PwC and the ACFE, one speaks only about detection. The studies investigate by means of surveys which are the most occurring means or methods that lead to fraud detection, or are believed to do so by the Chief Financial Officers. The following are the findings of both studies.

About the way fraud is detected, both studies of PwC and the ACFE stress the importance of tips. At the PwC study, no less than 43% of the fraud cases was detected by means of tips (whistle-blowing, internal tip-offs and external tip-offs). While the respondents of the ACFE study even reported a number of 46.2%. According to the ACFE report, an anonymous fraud hotline anticipates a lot of fraud damage. In the cases reviewed, organizations that had such hotlines, suffered a median loss of US$ 100.000, whereas organizations without hotlines had a median loss of US$ 250.000, presenting a reduction of 60% in fraud loss. Tips and hotlines are associated with a company's fraud culture. These numbers suggest that a company can detect and deter fraud by setting an appropriate tone at the top (and down). Another recent study, performed by Ernst&Young, also address the effectiveness of a Code of Conduct, expressing a company's culture (Ernst&Young, 2007). However, as a second best detection method, the company's control system comes forward. This is represented in several aspects, from internal audit to fraud risk management, but all together these methods contribute to fraud detection. At the PwC study, the corporate controls were responsible for the detection of 34% of the reported frauds. At the ACFE study, internal audit and internal control together revealed 39.4% of the reported cases. So also corporate control can have a measurable impact on detecting fraud after chance related means. The more control measures a company puts in place, the more incidents of fraud will be uncovered. Because of the domain field of this thesis (Accounting), we want to concentrate in this dissertation on the use of these corporate control methods. This selection is supported by the high impact controls have on fraud detection. Further, the methods of corporate control are more formal, better observable, less arbitrary and more measurable methods than corporate culture methods of detection.

Beware that all above mentioned suggestions concerning detection and prevention of fraud, concern *internal* fraud detection/prevention and further, are the results of non-academic research. In the next section, an overview of the academic literature concerning fraud detection and prevention is given.

1.2 Fraud Detection/Prevention Literature

For this literature review, articles on quantitative data analysis in the context of fraud detection and fraud prevention are searched. Literature on identifying factors that would facilitate fraud detection is not included. Also, the literature domain of internal control, which obviously is very closely related, is not reviewed as such. Only articles with a quantitative data analysis approach, as stated before, are included. Although the subject of fraud prevention is taken into account, almost all articles found address the problem of fraud detection. Also, mostly external fraud is apparently researched in the selected articles. This should not be interpreted as that internal fraud is not investigated as thoroughly. The fact that this literature does not come forward as strongly as external fraud literature, is the result of the asymmetry between both types of fraud concerning data analysis research.

To gain a clear view of the current situation of research, Table 1.1 is created. This will provide us with some insights of the implicitly followed methodology in current literature. The table provides us with the author(s) in alphabetical order, the application domain, whether it concerns internal or external fraud, whether the objective is fraud detection or prevention, and which technique is used. The information about the last column (Task) will be discussed later and is of no importance yet.

Concerning the techniques used, an intensively explored method are neural networks. The studies of Davey et al. (1996) and Hilas and Mastorocostas (2008) (telecommunications fraud), Dorronsoro et al. (1997) (credit card fraud), and Fanning and Cogger (1998), Green and Choi (1997) and Kirkos et al. (2007) (financial statement fraud) all use neural network technology for detecting fraud in different contexts. Lin et al. (2003) apply a fuzzy neural net, also in the domain of fraudulent financial reporting. Both Brause, Langsdorf, and Hepp (Brause et al.) and Estévez et al. (2006) use a combination of neural nets and rules. The latter use fuzzy rules, where the former use traditional association rules. Also He et al. (1997) apply neural networks: a multi-layer perceptron network in the supervised component of their study and Kohonen's self-organizing maps for the unsupervised part. (the terms *supervised* and *unsupervised* will be explained later). Like He et al. (1997) apply in their unsupervised part, Brockett et al. (1998) apply Kohonen's self-organizing feature maps (a form of neural network technology) to uncover phony claims in the domain of automobile insurance. This is also what Zaslavsky and Strizhak (2006) suggest later, in 2006, in a methodological paper to detect credit card fraud. Quah and Sriganesh (2008) follow this suggestion in an empirical paper on understanding spending patterns to decipher potential fraud cases. A Bayesian learning neural network is implemented for credit card fraud detection by Maes et al. (2002) (aside to an artificial neural network), for uncollectible telecommunications accounts (which is not always fraud) by Ezawa and Norton (1996), for financial statement fraud by Kirkos et al. (2007) and for automobile insurance fraud detection by Viaene et al. (2005) and Viaene et al. (2002).

In Viaene et al. (2005)'s field of automobile insurance fraud, Bermúdez et al. (2007) use an asymmetric or skewed logit link to fit a fraud database from the Spanish insurance market. Afterwards they develop Bayesian analysis of this model. In a related field Major and Riedinger (2002) present a tool for the detection of medical insurance fraud. They propose a hybrid knowledge/statistical-based system, where expert knowledge is integrated with statistical power. Another example of combining different techniques can be found in Fawcett and Provost (1997). A series of data mining techniques for the purpose of detecting cellular clone fraud is hereby used. Specifically, a rule-learning program to uncover indicators of fraudulent behavior from a large database of customer transactions is implemented. From the generated fraud rules, a selection has been made to apply in the form of monitors. This set of monitors profiles legitimate customer behavior and indicate anomalies. The outputs of the monitors, together with labels on an account's previous daily behavior, are used as training data for a simple Linear Threshold Unit (LTU). The LTU learns to combine evidence to generate high-confidence alarms. The method described above is an example of a supervised hybrid as supervised learning techniques are combined to improve results. In another work of Fawcett and Provost (1999), Activity Monitoring is introduced as a separate problem class within data mining with a unique framework. Fawcett and Provost (1999) demonstrate how to use this framework among other things for cellular phone fraud detection.

Another framework presented, for the detection of health care fraud, is a process-mining framework by Yang and Hwang (2006). The framework is based on the concept of *clinical pathways* where structure patterns are discovered and further analyzed.

The fuzzy expert systems are also experienced with in a couple of studies. So there are Derrig and Ostaszewski (1995), Deshmukh and Talluru (1998), Pathak et al. (2003), and Sánchez et al. (2008). The latter extract a set of fuzzy association rules from a data set containing genuine and fraudulent credit card transactions. These rules are compared with the criteria which risk analysts apply in their fraud analysis processes. The research is therefor difficult to categorize as 'detection', 'prevention' or both. We adopt the authors' own statement of contribution in both fraud detection and prevention. Derrig and Ostaszewski (1995) use fuzzy clustering and therefor apply a data mining technique performing a descriptive task, where the other techniques (but Sánchez et al. (2008)) perform a predictive task. (see further)

Stolfo et al. (2000) delivered some interesting work on intrusion detection. They provided a framework, MADAM ID, for Mining Audit Data for Automated Models for Intrusion Detection. Although intrusion detection is associated with fraud detection, this is a research area on its own and we do not extend our scope to this field. Next to MADAM ID, Stolfo et al. (2000) discuss the results of the JAM project. JAM stands for Java Agents for Meta-Learning. JAM provides an integrated meta-learning system for fraud detection that combines the collective knowledge acquired by individual local agents. In this particular case, individual knowledge of banks concerning credit card fraud is combined. Also Phua et al. (2004) apply a meta-learning approach, in order to detect fraud and not only intrusion. The authors base their concept on the science fiction novel *Minority Report* and compare the base classifiers with the novel's 'precogs'. The used classifiers are the naive Bayesian algorithm, C4.5 and backpropagation neural networks. Results from a publicly available automobile insurance fraud detection data set demonstrate that the stacking-bagging performs better in terms of performance as well as in terms of cost savings.

Cahill et al. (2000) design a fraud signature, based on data of fraudulent calls, to detect telecommunications fraud. For scoring a call for fraud its probability under the account signature is compared to its probability under a fraud signature. The fraud signature is updated sequentially, enabling event-driven fraud detection.

Rule-learning and decision tree analysis is also applied by different researchers, e.g. Hilas (2009), Kirkos et al. (2007), Fan (2004), Viaene et al. (2002), Bonchi et al. (1999) and Rosset et al. (1999). Viaene et al. (2002) actually apply different techniques in their work, from logistic regression, k-nearest neighbor, decision trees and Bayesian neural network to support vector machine, naive Bayes and tree-augmented naive Bayes. Also in Viaene et al. (2007), logistic regression is applied.

Link analysis comprehends a different approach. It relates known frauds to other individuals, using record linkage and social network methods (Wasserman and Faust, 1998). Cortes et al. (2002) find the solution to fraud detection in this field. The transactional data in the area of telecommunications fraud is represented by a graph where the nodes represent the transactors and the edges represent the interactions between pairs of transactors. Since nodes and edges appear and disappear from the graph through time, the considered graph is dynamic. Cortes et al. (2002) consider the subgraphs centered on all nodes to define communities of interest (COI). This method is inspired by the fact that frauds seldom work in isolation from each other.

To continue with link analysis, Kim and Kwon (2006) report on the Korean Insurance Fraud Recognition System that employs an unsupervised three-stage statistical and link analysis to identify presumably fraudulent claims. The government draws on this system to make decisions. The authors evaluate the system and offer recommendations for improvement.

Bolton and Hand (2001) are monitoring behavior over time by means of Peer Group Analysis. Peer Group Analysis detects individual objects that begin to behave in a way different from objects to which they had previously been similar. Another tool Bolton and Hand (2001) develop for behavioral fraud detection is Break Point Analysis. Unlike Peer Group Analysis, Break Point Analysis operates on the account level. A break point is an observation where anomalous behavior for a particular account is detected. Both the tools are applied on spending behavior in credit card accounts.

Also Murad and Pinkas (1999) focus on behavioral changes for the purpose of fraud detection and present three-level-profiling. As the Break Point Analysis from Bolton and Hand (2001), the three-level-profiling method operates at the account level and it points any significant deviation from an account's normal behavior as a potential fraud. In order to do this, 'normal' profiles are created (on three levels), based on data without fraudulent records. To test the method, the three-level-profiling is applied in the area of telecommunication fraud. In the same field, also Burge and Shawe-Taylor (2001) use behavior profiling for the purpose of fraud detection by using a recurrent neural network for prototyping calling behavior. Two time spans are considered at constructing the profiles, leading to a current behavior profile (CBP) and a behavior profile history (BPH) of each account. In a next step the Hellinger distance is used to compare the two probability distributions and to give a suspicion score on the calls.

A brief paper of Cox et al. (1997) combines human pattern recognition skills with automated data algorithms. In their work, information is presented visually by domain-specific interfaces. The idea is that the human visual system is dynamic and can easily adapt to ever-changing techniques used by frauds. On the other hand have machines the advantage of far greater computational capacity, suited for routine repetitive tasks.

A few last studies we would like to mention are those of Tsung et al. (2007), Brockett et al. (2002), Hoogs et al. (2007) and Juszczak et al. (2008). Tsung et al. (2007) apply manufacturing batch techniques to the field of fraud detection. They use the batch library method. Brockett et al. (2002) use a principal component analysis of RIDIT scores to classify claims for automobile bodily injury. Hoogs et al. (2007) present a genetic algorithm approach to detect financial statement fraud. They find that exceptional anomaly scores are valuable metrics for characterizing corporate financial behavior and that analyzing these scores over time represents an effective way of detecting potentially fraudulent behavior. Juszczak et al. (2008) at last apply many different classification techniques in a supervised two-class setting and a semi-supervised one-class setting in order to compare the performances of these techniques and settings.

If we summarize existing academic research by looking at Table 1.1, we arrive at the conclusion that merely all data analysis research is conducted in the field of external fraud. There clearly is a gap in the academic literature concerning internal fraud. Only six articles on internal fraud are found and they address only one kind of internal fraud: statement fraud. This is not even the number one internal fraud. Following the studies mentioned in Section 1.1.5 by PwC and ACFE, asset misappropriation, which is a form of transaction fraud, is the most prevalent kind of internal fraud. Transaction fraud is however no subject of existing research. Further is confirmed by Table 1.1 that the bulk of literature aims at providing a detection tool, only two articles incorporate the importance of prevention. As a last observation, one notices that all articles found apply data mining techniques. This is a remarkable divergence of the non-academic research, where internal control was pointed as an effective detection tool (after corporate cultural means). These findings are taken into consideration when delineating the research objective.

Table 1.1: Fraud detection/prevention literature overview

Author	**Application Domain**	**Internal/ External**	**Detection/ Prevention**	**Technique**	**Task**
Bermúdez et al. (2007)	Automobile Insurance Fraud	External	Detection	Skewed logit link and Bayesian analysis	Predictive
Bolton and Hand (2001)	Credit Card Fraud	External	Detection	Peer Group Analysis and Break Point Analysis	Predictive
Bonchi et al. (1999)	Fiscal Fraud	External	Detection	Decision Tree	Predictive
Brause, Langsdorf, and Hepp (Brause et al.)	Credit Card Fraud	External	Detection	Rules and Neural Network	Predictive
Brockett et al. (1998)	Automobile Insurance Fraud	External	Detection	Kohonen's Self-Organizing Map	Predictive
Brockett et al. (2002)	Automobile Insurance Fraud	External	Detection	Principal Component Analysis	Predictive
Burge and Shawe-Taylor (2001)	Telecommunications Fraud	External	Detection	Unsupervised Neural Network	Predictive
Cahill et al. (2000)	Telecommunication Fraud	External	Detection	Profiling by means of signatures	Predictive
Cortes et al. (2002)	Telecommunications Fraud	External	Detection	Dynamic Graphs	Predictive
Cox et al. (1997)	Telecommunications Fraud	External	Detection	Visual Data Mining	Descriptive
Davey et al. (1996)	Telecommunications Fraud	External	Detection	Neural Network	Predictive
Derrig and Ostaszewski (1995)	Automobile Insurance Fraud	External	Detection	Fuzzy Set Theory	Descriptive
Deshmukh and Talluru (1998)	Financial Statement Fraud	Internal	Detection	Rule-based Fuzzy Reasoning System	Predictive

Continued on next page

Author	Application Domain	Internal/ External	Detection/ Prevention	Technique	Task
Dorronsoro et al. (1997)	Credit Card Fraud	External	Detection	Neural Network	Predictive
Estévez et al. (2006)	Telecommunications Fraud	External	Detection and Prevention	Fuzzy Rules and Neural Network	Predictive
Ezawa and Norton (1996)	Uncollectible Telecommunications Accounts	External	Detection	Bayesian Neural Network	Predictive
Fan (2004)	Credit Card Fraud	External	Detection	Decision Tree	Predictive
Fanning and Cogger (1998)	Financial Statement Fraud	Internal	Detection	Neural Network	Predictive
Fawcett and Provost (1997)	Telecommunications Fraud	External	Detection	Rules, Monitors and Linear Threshold Unit	Predictive
Fawcett and Provost (1999)	Telecommunications Fraud	External	Detection	Activity Monitoring	Predictive
Green and Choi (1997)	Financial Statement Fraud	Internal	Detection	Neural Networks	Predictive
He et al. (1997)	Health Care Insurance Fraud	External	Detection	Neural Network	Predictive
He et al. (1997)	Health Care Insurance Fraud	External	Detection	Kohonen's Self-Organizing Map	Descriptive
Hilas and Mastorocostas (2008)	Telecommunications Fraud	External	Detection	Neural Network and Clustering	Predictive
Hilas (2009)	Telecommunications Fraud	External	Detection	If-then-else Rules	Predictive
Hoogs et al. (2007)	Financial Statement Fraud	Internal	Detection	A Genetic Algorithm Approach	Predictive
Juszczak et al. (2008)	Credit Card Fraud	External	Detection	Many different classification techniques	Predictive
Kim and Kwon (2006)	Insurance Fraud	External	Detection	Insurance Fraud Recognition System (Korea)	Predictive

Continued on next page

Author	Application Domain	Internal/ External	Detection/ Prevention	Technique	Task
Kirkos et al. (2007)	Financial Statement Fraud	Internal	Detection	Decision Tree, Neural Network and Bayesian Belief Network	Predictive
Lin et al. (2003)	Financial Statement Fraud	Internal	Detection	Fuzzy Neural Network	Predictive
Maes et al. (2002)	Credit Card Fraud	External	Detection	Neural Network and Bayesian Belief Network	Predictive
Major and Riedinger (2002)	Health Care Insurance Fraud	External	Detection	Electronic Fraud Detection (EFD)	Predictive
Murad and Pinkas (1999)	Telecommunications Fraud	External	Detection	Three Level Profiling	Predictive
Pathak et al. (2003)	Insurance Fraud	External	Detection	Fuzzy logic based expert system	Predictive
Phua et al. (2004)	Automobile Insurance Fraud	External	Detection	Meta-classifiers	Predictive
Quah and Sriganesh (2008)	Credit Card Fraud	External	Detection	Self-Organizing Maps	Descriptive
Rosset et al. (1999)	Telecommunications Fraud	External	Detection	Rules	Predictive
Sánchez et al. (2008)	Credit Card Fraud	External	Detection and Prevention	Fuzzy Rules	Descriptive
Stolfo et al. (2000)	Credit Card Fraud and Intrusion	External	Detection	Meta-classifiers	Predictive
Tsung et al. (2007)	Telecommunications Fraud	External	Detection	Batch Library Method	Predictive
Viaene et al. (2005)	Automobile Insurance Fraud	External	Detection	Bayesian Neural Network	Predictive
Viaene et al. (2002)	Automobile Insurance Fraud	External	Detection	Logistic Regression, k-Nearest Neighbor, Decision Tree, Bayesian Neural Network, SVM, Naive Bayes, and tree-augmented Naive Bayes	Predictive

Continued on next page

Author	**Application Domain**	**Internal/ External**	**Detection/ Prevention**	**Technique**	**Task**
Viaene et al. (2007)	Automobile Insurance Fraud	External	Detection	Logistic Regression	Predictive
Yang and Hwang (2006)	Health Care Insurance Fraud	External	Detection	Frequent Pattern Mining	Predictive

1.3 Research Objective

The lack of data analysis research on internal fraud in the academic field, is an opportunity to devote attention to. We decide to focus our research objective on internal fraud, with transaction fraud (not investigated at all) being a part of this.

That internal fraud is worth investigating, is already touched upon in section 1.1.5. The average financial damage to companies subjected to the PwC survey, was US$ 2.42 million per company and participants of the ACFE study estimated a loss of 7% of a company's annual revenues to fraud. Also Lynch and Gomaa (2003) draw attention to the susceptibility of organizations to fraudulent employee behavior. The authors state that with the integration of advanced information technology (IT) into business organizations, unintended risks and consequences can be introduced into the business environment.

Internal fraud has received a great deal of attention from interested parties like governments or non-profit institutions. The emergence of fraud into our economic world didn't go unnoticed. A US fraud standard (SAS 99) and an international counterpart (ISA 240) were created. Section 404 of the Sarbanes-Oxley act of 2002 also addresses this issue. Meanwhile, the CEO's of the International Audit Networks released a special report in November 2006. This report, issued by the six largest global audit networks, is released in the wake of corporate scandals. The authors of this report express their believe in fighting fraud, as they name it *"one of the six vital elements, necessary for capital market stability, efficiency and growth"*.[5] Unlike most fraud literature in academic fields, this report addresses internal fraud.

Another aspect a decision is made about, is the exact delineation of fraud research in this dissertation. One can focus on the factors that influence fraud, determinants that can be used to investigate fraud, fraud detection techniques, or fraud prevention mechanisms. This dissertation will aim at the combination of fraud detection and prevention, which will be referred to as 'fraud risk reduction'. This decision is corresponding with the ideas of Davia et al. (2000) and Bologna and Lindquist (1995), that fraud prevention and fraud detection should complement each other.

Based on the absence of a methodological framework to mitigate internal fraud in

[5] The remaining five elements concern investor needs for information, the alignment and support of the roles of various stake holders, the auditing profession, reporting and information quality.

the academic literature, the cost internal fraud nevertheless presents, and the clear interest the business environment shows, the research objective is to present and to apply a framework for internal fraud risk reduction. For this purpose, two courses are followed. We first have a look at what already exists in the business environment to prevent and detect internal fraud. Next, we turn to the methodology followed in the academic field with respect to external fraud. Particularly the use of data mining techniques are considered as a valuable contribution, since it has proven its value in mitigating external fraud. In the following chapter these two courses are explored, followed by the resulting framework for internal fraud risk reduction, the core of this dissertation.

Chapter 2

The IFR^2 Framework

The aim of this chapter is to develop a framework for reducing internal fraud risk. As already mentioned in Chapter 1 we approach this via two courses. We first have a look at what already exists in the business environment to prevent and detect internal fraud. On the other hand we examine the methodology followed in the academic field with respect to external fraud. These are the topics of the following sections. The gained insights result in the presentation of the framework for internal fraud risk reduction, which is entitled 'The IFR^2 Framework'.

2.1 Mitigating Internal Fraud in Practice: The Value of Internal Control

The studies of PwC and the ACFE mentioned before, reveal some information concerning the detection of internal fraud. The number one detection tool are tip-offs. This kind of detection method is linked with corporate culture, and not with corporate controls. The second best detection tool seems to be internal control, which will be highlighted in this section.

The (US) National Commission on Fraudulent Financial Reporting (or Treadway Commission) was formed in 1985. To study the causes of fraudulent reporting and make recommendations to reduce its incidence, the Treadway Commission issued a final report in 1987 with recommendations for auditors, public companies, regulators, and educators. This report re-emphasized the importance of internal control in reducing the incidence of fraudulent financial reporting and included a recommendation for all public companies to maintain internal controls. The Committee of Sponsoring Organizations of the Treadway Commission (COSO) [1] was formed to commission the Treadway Commission to perform its task. In response to this recommendation, COSO developed an internal control framework, issued in 1992 and entitled *Internal Control - Integrated Framework*. According to the COSO framework, internal control is defined as:

> *a process, effected by the entity's board of directors, management, and other personnel, designed to provide reasonable assurance regarding the achievement of objectives in the following categories:*
>
> - *Effectiveness and efficiency of operations*
> - *Reliability of financial reporting*
> - *Compliance with applicable laws and regulations*

Meanwhile, COSO issued in 2004 a revision of the *Internal Control - Integrated Framework* under the title of *Enterprise Risk Management Framework*, expanding on internal control to the broader subject of enterprise risk management. (Cosserat, 2004; Davia et al., 2000; Whittington and Pany, 1998)

If we look at the definition, it is clear why internal control is important as a protection against fraud. The achievement of the first category is to encounter transaction fraud, the second to encounter statement fraud and the third category achievement is to protect the organization against fraud for the company. Following this broad definition, internal control can both prevent and detect fraud. And although this definition is stemming from the foundation of the National Commission on Fraudulent *Financial Reporting*, also other classes of fraud than fraudulent financial reporting can be encountered. However, the definition is clear about its *reasonable* - not absolute - assurance regarding the objectives. We can conclude that internal control is a means to protect an organization against internal fraud, but given the raising prevalence of fraud it is still not sufficient as a stand-alone tool. Also the numbers provided by the PwC and ACFE surveys reveal that internal control comes off worse than chance means as a detection tool. However, these studies also emphasize the extra value of

[1] The sponsoring accounting organizations include the American Institute of Certified Public Accountants (AICPA), the American Accounting Association (AAA), the Financial Executives Institute (FEI), the Institute of Internal Auditors (IIA), and the Institute of Management Accountants (IMA).

well functioning internal control systems.

In its framework, COSO identifies the following five components as part of internal control:
- The control environment,
- The entity's risk assessment process,
- The information system,
- Control activities, and
- Monitoring of controls.

The control environment means the management's overall attitude, awareness and actions regarding internal control and its importance in the entity. The entity's risk assessment process comprehends the entity's process of identifying risks and apprehending an appropriate strategy towards these risks. Further, the entity should have access to an effective information system. This is an information system that guards the accurate reporting of transactions and provides a complete audit trail. Control activities are established by management to ensure that specific objectives are achieved. This need exists because of the information asymmetry already mentioned in the context of agency theory. The last component of internal control is monitoring. Employees need to know that non-compliance with controls is likely to be detected (deterrence effect). Monitoring controls also provides feedback concerning these controls. (Cosserat, 2004)

As becomes clear by these five components defined by COSO, internal control encompasses a wide variety of actions. In these components we see that aside from qualitative actions, like a control environment and a risk assessment process, also actions based on quantitative data are desired. The control activities for example may require some data analyzing. The techniques applied to this end are to be categorized as reporting tools, another category of data analyzing techniques than data mining. In the following section the difference between these two groups of data analyzing techniques is discussed.

2.2 Mitigating Internal Fraud in Academic Research: The Value of Data Mining

In this section, we turn to the contribution the academic field already delivered in the context of external fraud. The literature review in Chapter 1 pointed out that the academic community already investigated a broad range of data mining techniques for the purpose of detecting fraud and found some promising results. Although there is room left for a lot of other research in this area, we can conclude that there is a contributing value of applying the domain of data mining in a fraud context. For this reason, an introduction to this field is given. This also enhances to value the differences between our goal of *internal* fraud *risk reduction* versus the aim of most reviewed research in *external* fraud *detection*.

The current information age is overwhelmed by data. More and more information is stored in databases and turning these data into knowledge creates a demand for new, powerful tools. Data analysis techniques used before were primarily oriented toward extracting quantitative and statistical data characteristics. These techniques facilitate useful data interpretations and can help to get better insights into the processes behind the data. These interpretations and insights are the sought knowledge. Although the traditional data analysis techniques can indirectly lead us to knowledge, it is still created by human analysts. (Michalski et al., 1998) The current situation however needed a new way to deal with these never ending databases and new methods to analyze this huge amount of data. A new area came into being: Knowledge Discovery in Databases, also known as KDD. The process of KDD can be mapped out as in Figure 2.1, a representation based on Tan et al. (2006).

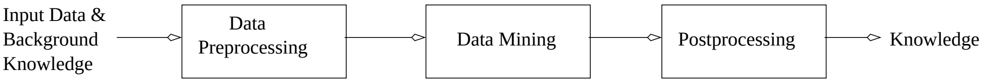


Figure 2.1: The process of knowledge discovery in databases

As we can see in this figure, an integral part of the process of KDD is data mining.

Together with KDD, data mining was born as a new research field. Data mining is a reaction to overcome the above limitations of data analyzing techniques used before. A data analysis system now has to be equipped with a substantial amount of background knowledge, and be able to perform reasoning tasks involving that knowledge and the data provided (Michalski et al., 1998). This is what data mining has an answer to. According to Witten and Frank (2000), data mining can be defined as

> *"...the process of discovering patterns in data. The process must be automatic or (more usually) semi-automatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic advantage. The data is invariably present in substantial quantities."*

In an effort to meet this goal, researchers have turned to ideas from different disciplines. The machine learning field for example is often mentioned in the same breath as data mining, since it has provided lots of input to data mining. However, data mining also relies on statistics, artificial intelligence, and pattern recognition. Data mining is a confluence of these disciplines.

With the coming of data mining as a new field of data analysis, data analyzing techniques can be divided into two groups: reporting techniques and data mining techniques. With reporting techniques we refer to the techniques used before, where quantitative and statistical data characteristics are extracted from data and human analysts turn this information into knowledge. Think for example at reports with some maximum, minimum and average numbers on sales or purchases. Also information on processes can be gathered and analyzed through reporting techniques. For example the quantity of time it takes between an incoming order and the delivery of the ordered goods. Here the information is focused on describing or learning more of the process itself, instead of the outcomes of the process. These are process reporting techniques. Reporting techniques are the types of techniques currently used in internal control settings. With data mining techniques we emphasize the (semi-)automatic process to discover meaningful patterns in large data sets. Especially the data mining characteristic of revealing latent knowledge is very typical and valuable. This characteristic comes forward in the fact that no hypotheses are needed to mine the data, as opposed to pure statistics or data reporting. This is the main reason why these techniques are selected in previous research for detecting external fraud. It is also this characteristic that makes the difference between the current software packages with business fraud risk solutions (such as ACL or IDEA) and the academic research described in the previous chapter. The software packages are merely interfaces for facilitating reporting.

An important step in applying data mining is that of data engineering. What data do we have, what kind of information does it capture and what knowledge do we want to extract from it? Depending on the field you (exa)mine, you have information about accounts. An account can involve several things, like a customer's account, an invoice, a calling account and so on. In fact we start from data about these accounts, which we call account data. For example, for a customer's account, what is the name of the customer, where does he live, what is his telephone number, when did he become a customer and so on. We do not only have account data, we also have operational information about an account. This kind of data describes the behavior of an account, like what was bought on an account, when was something bought, were there any reductions and so on. So actually we have two kinds of information available: account data and operational data on the account. A data mining approach links this information and attempts to alter technical data into behavior since the purpose of a data mining approach is to discover patterns in data.

There are many techniques the field of data mining encompasses, like K-means clustering, decision trees, neural networks etc. These techniques serve different tasks, like for example classification, clustering, and anomaly detection. Mainly, data mining tasks can be divided in two subgroups: predictive tasks and descriptive tasks. With predictive tasks, the objective is to predict the value of one attribute, based on the values of other attributes. This is what classification techniques pursue. Predictive tasks make a prediction for every observation. Descriptive tasks however, do not pronounce upon every observation, but describe the data set as a whole. It aims to describe the underlying relationships in the data set. Examples of descriptive tasks are pattern recognition, anomaly detection, and correlations. (Tan et al., 2006)

In Table 1.1 an additional column is provided, stating what kind of task is used in a particular article. In the case of academic fraud detection literature, it appears that mainly predictive tasks are executed. Many different techniques serve this end. The class to be predicted in this context is the label 'fraudulent'/'non-fraudulent'.

Aside from grouping data mining applications based on their task (predictive versus descriptive), there is yet another dimension to classify learning algorithms. Based on the input data, there are two categories of learning: supervised and unsupervised learning. In supervised learning, the class to be learned is present in the data set. In the fraud detection problem, this translates in a data set containing examples of both fraudulent and non-fraudulent records. This means that all the records available are labeled as 'fraudulent' or 'non-fraudulent'. After building a model using these training data, new cases can be classified as fraudulent or non-fraudulent. Of course, one needs to be confident about the true classes of the training data, as this is the foundation of the model. Another practical issue is the availability of such information. Furthermore, this method is only able to detect frauds of a type which has previously occurred. In contrast, unsupervised methods don't make use of labeled records. These methods seek for accounts, customers, suppliers, etc. that behave 'unusual' in order to output suspicion scores, rules or visual anomalies, depending on the method. (Bolton and Hand, 2002)

Whether supervised or unsupervised methods are used, note that the output gives only an indication of fraud likelihood. No stand alone statistical analysis can assure that a particular object is a fraudulent one. It can only indicate that this object is more likely to be fraudulent than other objects.

Mainly supervised data is used in the external fraud detection literature. With Bolton and Hand (2001), Murad and Pinkas (1999), Burge and Shawe-Taylor (2001), Brockett et al. (2002), Kim and Kwon (2006), Cox et al. (1997) and Quah and Sriganesh (2008), the most important studies concerning unsupervised learning in fraud detection are quoted. Although this list may not be exhaustive, it is clear that research in unsupervised learning with respect to fraud detection is due for catching up. This is also a possible explanation for the 'transaction fraud gap' in the literature. There is no supervised data available on this kind of fraud. The only internal fraud with supervised data available is statement fraud, not coincidentally the only kind of internal fraud investigated in the academic literature. We have to take this difference into consideration when creating our methodology for internal fraud.

2.3 The IFR^2 Framework

Starting from the insights of the previous two sections, we develop the IFR^2 Framework in this section.

Internal fraud is currently dealt with by internal control. Internal control is embedded in a well elaborated framework, established by the COSO and encompasses a wide variety of tasks and settings. Next to a qualitative approach (like for example creating a control environment), data analyzing is required. It is at this point an opportunity lies to combine academic research with practical insights. Data mining tools are not implemented in the internal control framework. However, academic research already revealed that data mining can be of surplus value in detecting fraud in an external fraud context. The IFR^2 Framework advises to incorporate the data mining field as a way to implement the fourth component of COSO, control activities.

However, because there are elements of distinction between found academic research and our aim, we cannot just copy existing methods of working. Two major differences between our objective and existing work is that we 1) focus on internal fraud which typically involves unsupervised data, and 2) focus on fraud risk reduction instead of fraud detection. This includes both fraud prevention and fraud detection, as opposed to existing literature which only investigates the use of data mining for fraud detection. These differences will have their effect on our IFR^2 Framework, which will differ from the methodology (although never explicitly registered!) followed in existing literature. The IFR^2 Framework is presented in Figure 2.2.

The IFR^2 Framework starts with **selecting a business process with an advanced IT integration**. An organization should select a business process which it thinks is worthwhile investigating. This selection can be motivated by different aspects: a business process that has a great cash flow, one that is quite unstructured, one that is known for misuses, or one that the business has no feeling with and wants to learn more about. Also the implementation of advanced IT, according to Lynch and Gomaa (2003), is a breeding ground for employee fraud.[2] So selecting a business process with an advanced IT integration, as opposed to a paper based environment, is a good starting point to encounter this stream of frauds.

After the selection of an appropriate business process, **data has to be collected, manipulated and enriched** for further processing. This is comparable to the step ”Data preparation” in Chien and Chen (2008)’s framework for personnel selection. The manipulation of data refers to the cleaning of data, merging connected data, transforming data into interpretable attributes and dealing with missing values. Enrichment stands for creating extra attributes like for instance some ratios. Although background knowledge may be required for executing this step, these are mainly tech-

[2] The term of advanced IT integration refers to the ever presence of IT in a business process of a company.

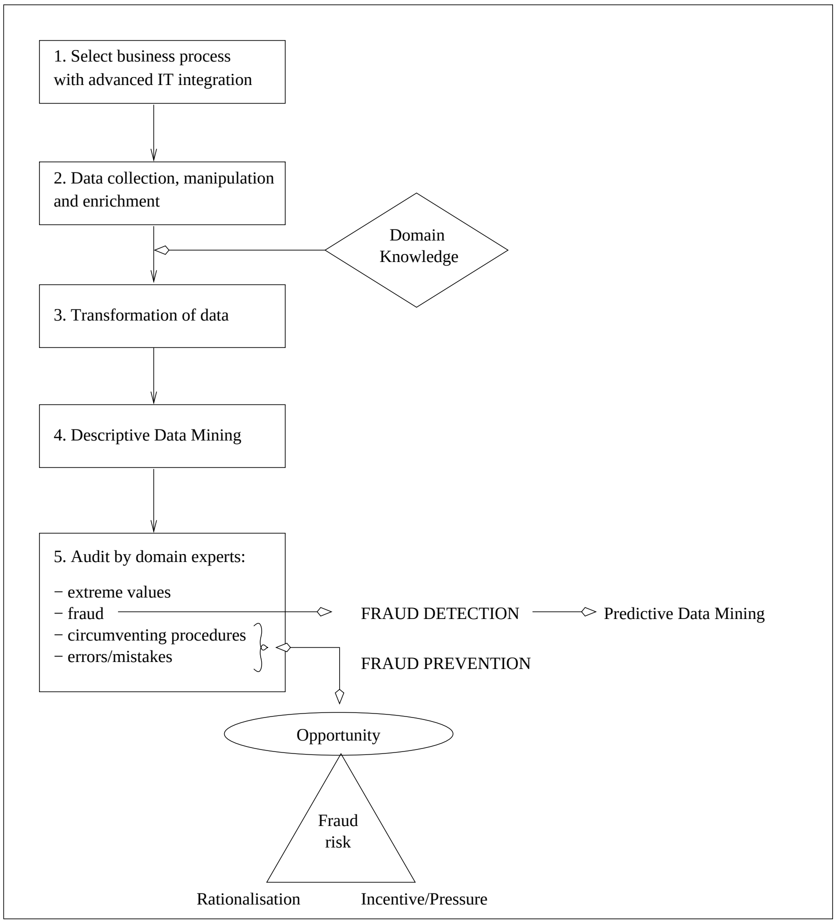

1. Select business process with advanced IT integration
2. Data collection, manipulation and enrichment
Domain Knowledge
3. Transformation of data
4. Descriptive Data Mining
5. Audit by domain experts:
− extreme values
− fraud
− circumventing procedures
− errors/mistakes
FRAUD DETECTION
Predictive Data Mining
FRAUD PREVENTION
Opportunity
Fraud risk
Rationalisation
Incentive/Pressure

Figure 2.2: The IFR^2 Framework

nical transactions in that they still present operational data.

During the third step, **transformation of the data**, the operational data will be translated into behavioral data. This translation builds - even more than the second step - upon domain knowledge and is not just a technical transformation.

The core of the framework is then to apply a **descriptive data mining** approach for getting more insights in this behavioral data. This is where the IFR^2 Framework remarkably differs from the followed methodology in the existing literature. In academic literature, almost all research applies a data mining technique with a predictive task. The explanation for the IFR^2 approach is twofold. Existing work predicts whether an observation is fraudulent or not. This can be explained by their focus on fraud detection. We however broaden our intentions, and are interested in all information captured in the data that helps us reducing the fraud risk, and not only the class 'fraudulent/legal'. In order to retrieve more information and patterns in data, a descriptive data mining approach has to be pursued.

Another characteristic of internal fraud risk reduction is the presence of unsupervised data sets, liable to this stream of research. There are almost no supervised data sets available in the context of internal fraud. This fact also accounts for the use of descriptive data mining instead of predictive data mining. An advantage of the use of descriptive data mining techniques is that it is easier to apply on unsupervised data. Thus for overcoming the exclusion of types of fraud where supervised data is difficult to obtain, the use of descriptive data mining techniques is recommended.

The core of this methodology -to use descriptive data mining- is also motivated by the higher intrinsic value a description of the data set under investigation provides than just a prediction of fraudulent versus legal. A description of the data set as a whole can bring insights to light, that were not clear before. All extra insights an analyst can gain, are valuable to better understand what is going on. This will eventually lead to a better position to mitigate internal fraud. When one only focuses on predicting the fraud class, one is not open minded enough to notice other interesting patterns. Association rules, clustering and anomaly detection are appropriate candidates for describing the data set. These can ultimately lead to observations or outliers, seeming interesting to take a closer look at. This is what happens in the fifth step of our methodology.

The fifth step is the **audit of interesting observations by domain experts**. The descriptives should provide the researchers a recognizable pattern of procedures of the selected business process. In addition some other patterns of minor groups of observations in the data can arise, interesting to have a closer look at. By auditing these observations, one can acquire new insights in the business process. As a general rule, one will always select outliers or extreme values to take a closer look at. Observations defined as outlier can normally be brought back to one of the following four cases: the observation is a (regular) extreme value, the observation is fraudulent, the observation is the result of circumventing procedures or it is simply a mistake. The regular observations will not draw our attention.

Observations defined as an outlier because they contain extreme values -but very natural when looked into- are not of interest for our purpose. (Think for example at the purchase of a mainframe at the same department as the purchases of CDs.) Nevertheless, they can occur. The other three categories (fraud, circumventing procedures and mistakes) on the other hand are of interest. If a fraudulent observation comes to our attention as an outlier, this is part of fraud detection. A fraud case can be interesting for adjusting current practice in the business process. If enough similar fraud cases are uncovered, a supervised fraud detection method can be elaborated for this specific fraud, based on a new data set. In this particular case, one can find well elaborated and tested methods in the existing literature. At this stage of investigating, **predictive data mining** tasks are recommended to search specifically for this type of fraud. The other two categories which can be at the origin of an outlier, circumventing procedures and making mistakes, are important in the light of fraud prevention. By making a mistake and realizing nobody notices or by circumventing procedures, a window of opportunity to commit fraud can develop. Opportunity, aside from rationalization and incentive or pressure, is one of the three elements of the fraud triangle[3] (see Section 1.1.3). The fraud triangle has been formally adopted by the auditing profession as part of the Statement on Auditing Standards No. 99. Also according to Albrecht et al.'s 'fraud scale' and even according to Hollinger and Park's theory, opportunity is an element of influence on fraud risk. Being able to select those cases where procedures are circumvented or mistakes are made, is an important contribution to taking away this opportunity and hence to prevent future fraud. The way in which this is dealt with, is up to the company. Internal controls can be adapted, persons can be called to account, procedures can be rewritten or other measures can be taken. This follow-up is not part of this framework anymore.

Tennyson and Salsas-Forn (2002) show that claims auditing, in the field of automobile insurance fraud, works as fraud detection and as fraud deterrence (a way of preventing) as well. This proves the value of the fifth step of our methodology.

[3] We refer to the general fraud triangle instead of a more specified theory like 'Broken Trust' or 'American Dream', because these latter are not applicable to every kind of internal fraud.

2.4 Conclusion

The IFR^2 Framework presented in this chapter has three major contributions. Firstly, it provides a guidance in internal fraud risk research. This was not present yet in the academic literature since almost all research was conducted on external fraud. Secondly, the core of the IFR^2 Framework is to apply a descriptive data mining approach. This is innovative in academic research in the context of fraud and data mining, and it is on top of this of value for organizations. Organizations mostly implement an internal control environment to mitigate internal fraud. This environment, mostly based on the COSO framework, does not imply the field of data mining. Extending the current internal control settings in organizations with the advantages of data mining, will deliver additional insights to reduce internal fraud risk. And thirdly, the IFR^2 Framework is not focusing on fraud detection only, but on detection and prevention. Hence, fraud risk is reduced instead of only detecting fraud when it already took place. The idea is to work more pro-actively.

Chapter 3

Business Process Selection

The IFR^2 Framework introduced in Chapter 2 will be implemented in 'Epsilon'. This company, which chooses to stay anonymous in this study, is an international financial services provider, ranked in the top 20 of European financial institutions. In this chapter, the first step of the IFR^2 Framework is executed: selection of a business process with advanced IT integration. As described in this chapter, the procurement process is selected as the business process subjected to this study. A process analysis and risk assessment are conducted, leading to the identification of some "hot spots' within this process.

This chapter first provides an overview of the most important business processes in a company. Next, the selection of the procurement business process is explained. In order to gain insights in the procurement process followed by Epsilon, a process analysis and risk assessment are conducted. The general outlines of a process map are described before the actual procurement process is expounded.

3.1 Company Business Processes

The implementation of business processes is, for an organization, a general method to achieve its objectives. In order to select an appropriate business process for this study, one needs to identify the business processes in place in organizations. In order to do this, the core activities, which are translated to processes, need to be known. An answer on this question is provided by Michael Porter. In his book, Competitive Advantage, Porter introduced the concept of generic value chains. (Porter, 1985) Such a generic value chain holds a sequence of core activities that a wide range of companies show. From an organizational perspective, these core activities are translated into core business processes, to be found in most companies. Accordingly, we will discuss these activities as generic business processes. (Knechel, 2001)

The value chain activities are to be divided in primary activities and support activities. The **primary activities** are:

- **Inbound logistics**,
  These activities refer to receiving input and resources and storing them until needed for the value creating process.
- **Operations**,
  These activities contain the core business of the organization, the transformation of inputs to goods to be sold.
- **Outbound logistics**,
  These activities concern the trajectory from production to customer, including warehousing and distributing.
- **Marketing and sales**,
  These activities aim to create demand for the produced or delivered good or service.
- **Service**.
  These activities represent the customer support and service after delivery of the goods or services, including warranty and claims handling.

These primary activities will lead to a higher company's margin, when executed efficiently.

The primary activities are supported by support activities, also contributing to a higher margin when executed efficiently. The **support activities** are:

- **Firm infrastructure**,
  These activities concern, amongst other, general management, accounting, finance, control systems, public affairs, ... often referred to as 'corporate overhead'.

- **Human resource management**,
  These activities concern personnel issues, like hiring employees, training, and compensation.

- **Technology development**,
  These activities aim to facilitate or to turn the core activities more efficiently.

- **Procurement**.
  These activities concern purchases of raw material, supplies, spare materials, buildings, equipment, etc.

Knowing the core business processes in an organization, a selection has to be made. The selection of a business process as subject of investigation can be based on several different aspects. One aspect that is suggested in the IFR^2 Framework is the degree of IT integration. Business processes without IT integration are for that reason not withheld. At Epsilon, all business processes are integrating a considerable degree of IT. Accordingly, the aspect of IT integration is not a good selection criterium in this case. Another criterium to select the business process with a considerable fraud risk, is to combine the occurrence probability of fraud and the associated cost of undetected frauds at a particular business process. A risk assessment of all business processes to end up with an objective mathematical answer on previous question would take another dissertation. In the case of Epsilon, we called in the help of the Investigations department (i.e. part of firm infrastructure). This department handles all fraud cases, internal and external, of Epsilon and works closely with the department of internal control and internal audit. Accordingly one can assume the Investigations department has a clear view on which business process contains which risks.

According to the Investigations department, a lot of money is involved in the procurement business process (Epsilon purchases annually for around 1.4 billion euros). Interestingly, unlike all other business processes, this business process never files a fraud at Investigations. Because this business process is believed to have an equal probability on fraud occurrence as any other business process, this lack of fraud files is seen as an indicator of fraud cases that stay undetected, more than they do in other processes. On top of this, because of the large amounts of money involved, undetected frauds can represent considerable costs. As a result, the procurement business process was found suited for the investigation of internal fraud risk reduction. In order to perform the second step of the IFR^2 Framework, i.e. data collection about this selected process, some background information about the process needs to be acquired. This is done by means of a process analysis and risk assessment. In the next section, a general, theoretical outline of process analysis is given and in the next following section the information acquired by this analysis at Epsilon is presented. A risk assessment at Epsilon is described in the final section. Performing a process analysis and risk assessment is not representing a step in the IFR^2 Framework, because it is assumed that organizations which follow this framework are familiar with the core outlines of the selected business process. So when researchers and practitioners are collaborating closely, this step may be discarded. However, so as to provide all readers -including the external researchers, who may have to include this extra step-with enough background information, this hidden step is discussed thoroughly.

3.2 Performing a Process Analysis

When analyzing a process, a great deal of information needs to be acquired. A systematic approach for gathering and documenting process details is essential in this analysis. Accordingly, a *process map* can be used. A process map consists out of four components to document details about a business process: 1) process objectives, 2) process activities, 3) information flows, and 4) the accounting impact. (Knechel, 2001)

The process objectives specify the purpose of the business process. The objectives can vary in number and importance between and in processes. Not all process objectives within one process are equally important. The second component of process analysis, identifying the constituting process activities, is the most self-evident component. While the process objectives inform about the strategic purpose, this second component translates the first one into operational activities. The third kind of information concerns process data streams. What information feeds are used during the business process and what information is generated? Also the type of information flow is important, as is the reliability of the information system. An electronic data flow is a completely different one than a manual data flow. This kind of information needs to be gathered too. The final component refers to collecting data about the accounting impact of the business process. (Knechel, 2001)

A process map, as explained in the previous paragraph, is made for the procurement process at Epsilon. For collecting the desired information, various methods are applied. Executive officers are interviewed, employees at various departments are questioned and observed during their job, internal user guidelines of the Enterprise Resource Planning (ERP) system are consulted, along with reports of internal audit and of risk management. The information gathered during the process analysis is not reported in the same fashion as a process map though. For reasons of comprehension for the external researcher, a full text and description of the process is given, rather than a bullet list addressing the four components. All four types of information are however covered and presented in Section 3.3, where the procurement process of Epsilon is described.

3.3 Procurement Process at Epsilon

The overall purchasing process can be visualized as in Figure 3.1. A purchase is triggered by an internal request. The request will be directed to the Subject Expert of the desired object. This Expert fills out a Purchase Request Form that is handed over to the Tactical Purchasing Cell (TPC) in charge. There are about 140 TPC's within Epsilon, scattered over a few business lines. Each TPC is in charge of the purchases in a well specified domain. Whenever a purchase is of strategic nature, the Subject Expert and the TPC enter into consultation with the Strategic Purchasing Organization. After consultation, if any, the TPC places an order with the supplier, the goods or services are delivered and an invoice will be sent to Epsilon. These invoices are handled at the Invoice department.

Purchases can also be created without following this formal process and hence an invoice will arrive at the Invoice department without passing through the described process. However, in all cases Invoice will be involved in the purchasing process. Epsilon prefers most of the purchases to pass also through the TPC's, although they realize that this is not always possible. The Strategic Purchasing Organization at last, doesn't interfere so commonly, as was already pointed out. In what follows, the SPO, TPC's and Invoice will be discussed.

3.3.1 Strategic Purchasing Organization

Although the Strategic Purchasing Organization is not always involved in the purchasing process, it is best to start with the description of this department. For whenever the Strategic Purchasing Organization comes into play, it generates input for the

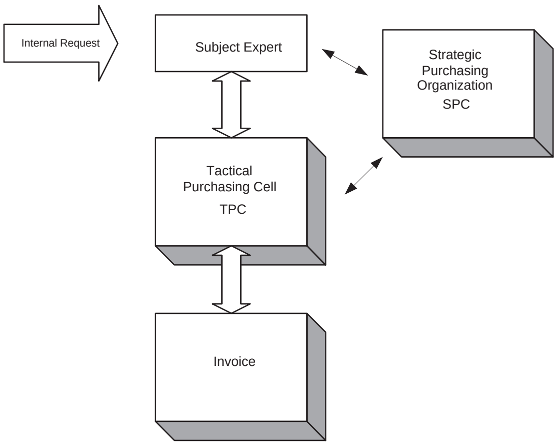


Figure 3.1: Overall Purchasing Process

other departments. The Strategic Purchasing Organization, or SPO, interferes when purchases encompass large amounts of money. This can be caused by a once-only purchase (e.g. a new mainframe) or by a contract that implies recurrent costs in the future (e.g. cleaning or catering). In both cases a close collaboration between the Subject Expert, the TPC and the SPO is desirable. The aim is to provide guidance in selecting an appropriate supplier that best satisfies the occurring need. A large part of this guidance exists in negotiating with several potential suppliers. After negotiating a contract, an Outline Agreement is made. This is the official standard form for contracts at Epsilon. The functioning of the SPO in this process is explained more in detail in the following sections.

Four Leverages of Spend Management

The aim of the SPO is to enforce Spend Management. This has been attained by paying attention to four leverages:

- Demand Management,
- Value Engineering,
- Supplier Management, and
- Compliance.

Demand Management questions the formulated need for new purchases. ”What is it they need?” ”Is the proposed purchase item the right solution, or are other items better fit to satisfy the occurring needs?” Value Engineering pays attention to the proportion of quality over price. Supplier Management aims at finding the ’right’ suppliers to join forces with. Compliance is the fourth leverage, but nevertheless very important. ”Are contracts complied with?” ”Do suppliers deliver as they said they would?” But Compliance does not only involve suppliers. ”Do employees of Epsilon make use of the contracts negotiated?”

The SPO works with separate projects. Each project follows a trajectory. Within these trajectories, the four leverages are paid attention to.

Four Spend Domains

The SPO supports the TPC’s and Subject Experts in all possible contracts that are large enough. The subjects of these contracts vary from paper and pencils over catering to maintenance of buildings. To get some kind of overview, the subjects of the supported projects are categorized in four spend domains. Each domain has its Spend Manager who is responsible for the purchases in his or her domain. These are:

- IT,
- Facilities,
- Human Resources et al.,
- Catalogue.

IT holds all purchases concerning information technology. Within Facilities fall all purchases related to supplies, e.g. furnishing and office supplies. Human Resources et al. categorizes all expenses linked with human resources and marketing, like hiring and training personnel, publicity and media. The contracts categorized in Catalogue all deal with purchases marked by their high volume but low price level per unit. These objects are generally used by every department of Epsilon, like paper for example.

Annual Plan of Projects and Project Teams

Each Spend Manager draws up an annual plan of purchases. This plan is based on internal requests as well as on domain expertise. The plan of purchases is converted into a plan of projects, by which each project follows a trajectory. As already mentioned, this trajectory watches over the four spend leverages.

Although a trajectory is developed to cover the four spend leverages, the main concern of the Spend Manager are the first and the fourth leverage, Demand Management and Compliance. During the trajectory, the Spend Manager gets help from the Project Office. This Project Office composes each time a Project Team of different people. In general the Project Team is composed of a Spend Project Coordinator, a Buyer and a Contract Engineer, each with his own function within the team. Next to these people, coming from the SPO itself, a Project Leader and a Subject Expert apply at the Project Team. The Project Leader and the Subject Expert are coming from the business line involved with the project.

Rules of Engagement

Since the SPO cannot foresee all large purchases within the four domains, there are rules of engagement for when to enlist the help of the SPO. In accordance with these rules, a business line has to notify the SPO when a purchase between $125,000$ and $250,000$ euros is planned. A purchase of a larger amount than $250,000$ euros is accompanied by the official help of the SPO.

However, these rules of engagement have two major problems. In the first place, these rules are not yet officially communicated among employees in Epsilon. Secondly, it is not clear what is meant with the amounts. Is it the height of one purchase, or the sum of all purchases with one supplier a year, or half a year. . . ? This shortcoming on communication is important in the light of risk assessment.

Ethical Code of Conduct

Every employee of Epsilon involved in purchasing is obliged to sign the ethical code of conduct. This code treats Epsilon's point of view with regard to its suppliers. Officially every employee of Epsilon involved in purchasing is supposed to have signed this code, but in practice this is confined to the employees of the SPO.

Compliance

A few times a year (or once a year, this depends on the Spend Manager), the Spend Manager analyzes the purchases in his or her domain. This analysis is done by means of reports drawn from a Business Objects Universe. This is a data warehouse built on the SAP system used in Epsilon. By analyzing these reports the Spend Manager can get an overview on how many purchases are made within his domain, how many of them were linked with a negotiated agreement, which business line consumes most, which business line purchases a lot by itself instead of in cooperation with the SPO etc. The analysis of these reports is in the light of the fourth leverage, Compliance. As mentioned before, this is an important part of the Spend Manager's function, and hence of the SPO.

3.3.2 Tactical Purchasing Cells

Epsilon counts about 140 Tactical Purchasing Cells, or TPC's. These are scattered over a few business lines, but most of them are concentrated in the business lines Facility and IT. TPC's not embedded in one of these two business lines are responsible for very specific purchases, e.g. the business line 'Design' has its own TPC's. The task of the TPC's is to enter the Outline Agreements, negotiated by the SPO, and Purchasing Orders into the SAP system. Only people working at a TPC have a user profile that is authorized to create a Purchasing Order or an Outline Agreement. Once an Outline Agreement or Framework Contract is entered into the system, a TPC can later attach Purchase Orders to it. These Purchasing Orders, whether or not attached to an Outline Agreement, cause orders with suppliers, which in turn cause invoices. These invoices will not arrive at the TPC's, but will be handled at Invoice. Invoice however may make an appeal to the TPC's while doing this job. Nevertheless, providing Invoice of information is a minor part of the job of the TPC's. So as to give, a more detailed description of the functioning of the TPC's. The visualization of this process is provided in Figure 3.2.

An internal request is received at the TPC's. Officially this occurs on the basis of a Purchasing Request Form. However, this is not common in practice. Mostly an e-mail is sent by the Subject Expert to the correct TPC in order to make a Purchasing

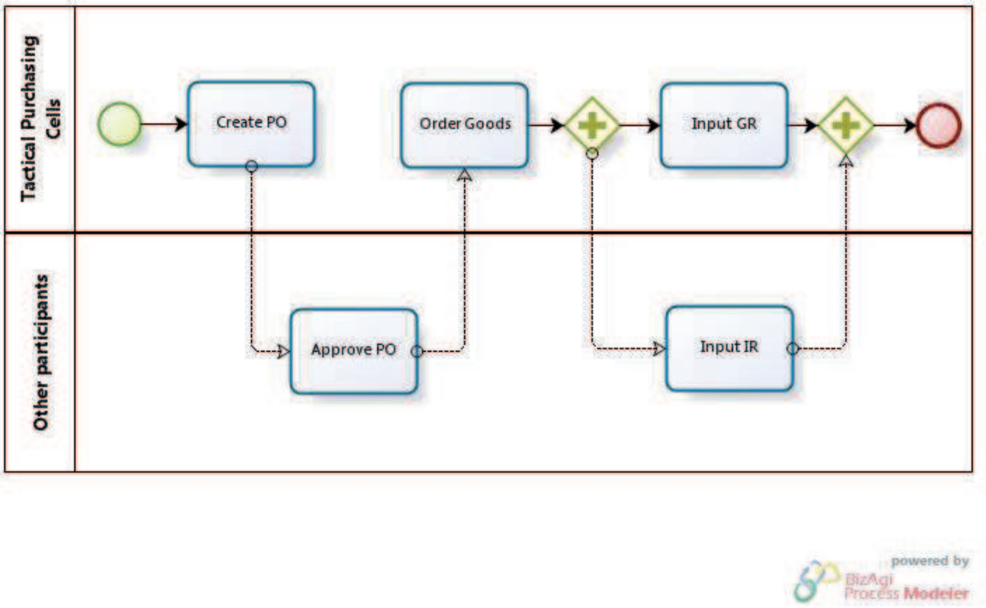


Figure 3.2: Tactical Purchasing Cell's activities

Order, or PO. Each person at a TPC knows the Subject Expert they are supposed to get orders from. Internal requests directly from Epsilon personnel are returned to sender. Everybody has to pass its requests on to the Subject Expert.

Whenever a request received, the TPC looks for Outline Agreements concerning this subject. If an existing Outline Agreement is found, a PO is attached to this Agreement. The benefit of this method of working is the point-and-click system provided to create a PO. The requested purchasing items only have to be clicked on in a list and the right prices appear. Also no choice has to be made with regards to the supplier. If no Outline Agreement is on hand, a stand-alone PO is created. The creation of a PO, as the creation of an Outline Agreement, passes in the SAP environment.

The information entered into the system when creating a PO is miscellaneous. The supplier that will deliver the items is mentioned, as are its coordinates and the expected date of delivery. The description of the purchasing items goes into details:

what, how many, price each, which cost center and which cost element. Also the person who creates the PO, the date of creation, the purchasing group[1] etc. are saved. In addition each PO is saved with a PO-number.

Whenever a PO is created, the system will start a work flow. This work flow creates a work item in the inbox of two persons who have to approve the purchase. First a work item is created in one person's inbox, next in the other one's inbox, after the first person has approved. This approval is done electronically in the SAP environment. When the second person approves, the PO is 'released'. Only after this electronic release, the TPC can hand over the PO to the supplier. The persons responsible for approval are captured in tables of authorization. Each business line makes such a table of authorization. These names vary with type of purchase item and with amount of purchase.

After release of the PO, the purchase items are ordered at the supplier. This is mostly done by means of an e-mail. Simultaneously, a Goods Receipt form is handed over to the Subject Expert. On this form the details of the order are stipulated. When the order arrives later on, the Subject Expert signs this Goods Receipt form and returns it to the TPC. This form is up till now still on paper. The next task of the TPC is to enter this document in the system. The person at the TPC opens the PO that triggered the delivery and fills out the accompanying electronic Goods Receipt. From here, the process is in hands of other departments. Sometimes in the last phase, when invoices are handled, the TPC is called in for information on some items. Otherwise, this was their last task.

As already mentioned, in Figure 3.2 one can see the tactical purchasing process, with the role of the TPC's in it. However, one should notice that not all tactical purchases take place this way. Several purchases happen without a PO and hence do not pass this process. For the TPC's this does not make much of a difference. Only at Invoice these purchases have a different handling. More explanation on this topic is given in Section 3.3.3.

[1] The purchasing group is a division based on cost center or on cost element, this is not always the same and hence not always very useful.

3.3.3 Invoice

In 3.3.1 and 3.3.2, the method of negotiating contracts or ordering purchase items with suppliers is discussed. In this section the last part of a purchase is described, the handling of invoices. This is the responsibility of the Invoice department, referred to as Invoice. Expense claims, company credit cards, and incoming and outgoing invoices are all handled at Invoice. Since expense claims, company credit cards, and incoming invoices are involved in the purchasing activity in the broadest sense of the word, these are shortly discussed. Outgoing invoices are left out of this discussion, because these have nothing to do with the purchasing activity of Epsilon. This process is an example of firm infrastructure activities.

Expense Claims

Besides an automatic handling of the expense claims, Invoice checks manually the included vouchers, if required. For expenses that do not require vouchers, Invoice checks manually for inconsistencies, e.g. a business lunch at Sunday.

Company Credit Cards

Management personnel at Epsilon all have a company credit card. This card can be used for business as well as for personal purposes. For the business part, vouchers are required to prove expenses. The part of the bill that has not been proven is subsequently assumed to be for personal consumption. This amount then is taken off of the monthly salary.

Since March 2005 employees of Epsilon can request and justify their accounts through the intranet application of Epsilon. At Invoice this justification is matched with the manually provided vouchers. This match covers two aspects: presence and correctness. Correctness alludes mainly to the right percentage of tax deduction.

Incoming Invoices

Epsilon handles about 170,000 incoming invoices a year. A project of automatization is started and introduced per business line. According to this project, incoming invoices are scanned before handled. This project is called Workflow&Scanning. The same project is responsible for the work flow items automatically delivered in the inbox, which is described in the tactical purchasing process.

Whether or not a business line already started with the project of automatization, does not make much of a difference for the process of invoice handling. Yet another feature is of more importance, namely whether the invoice corresponds to a purchase ordered with or without a PO. Whenever a purchase is triggered by an official PO of Epsilon, the approval is already given in the process of creating the PO. At the stadium of paying the corresponding invoice, approval is redundant. If the purchase however is not triggered by an official PO of Epsilon and one has not passed the entire process described in Section 3.3.2, approval still has to be given before paying the invoice. This difference in handling groups invoices in two kinds of invoice: PO-invoices and nonPO-invoices. The following paragraphs describe the handling of both kinds of invoices.

PO-invoices

Invoice receives the invoice. This can be physically on paper or as a work item within the SAP environment, depending on the status of Workflow&Scanning. Suppliers have been asked to mention the PO-number as a reference, so that employees at Invoice can link the invoice at the right PO. Once the invoice is entered into the SAP system, two possibilities can occur. Or the invoice will be posted (normal case), or the invoice will be parked (if something is missing or does not match). When parking occurs, the invoice is sent to the corresponding TPC. This cell has to complete the missing fields in the system. These are the fields Invoice was not able to fill out. Once Invoice has all the required information, the invoice will be posted.

When posted, the SAP system runs an internal control to match the Goods Receipt, entered by the TPC's, to the invoice, entered by Invoice. When there is a match, the invoice will not be blocked and payment will automatically follow. However, when there is no match, the invoice will be blocked, and the TPC has to solve this problem. This can be done in three ways:
- Change PO (this can trigger a new approval procedure),
- Request a credit note,
- Enter an additional Goods Receipt.

As the problem is solved, the invoice will be paid.

NonPO-invoices

As mentioned before, handling PO-invoices does not include an approval procedure. The approval has already been given at the stadium of creating the PO, as described in Section 3.3.2. A nonPO-invoice is different from a PO-invoice in its stadium of approval. Since a nonPO-invoice arrives at Epsilon without notice, the approval still has to be given at this late moment.

The receival of a nonPO-invoice is the same as with a PO-invoice. However, there is an additional point of interest: does the business line responsible for the invoice work with SAP or not? Within Epsilon, not all business lines work with SAP. If this is the case, the approval of an invoice has to take place by means of physical signatures. An approved invoice always carries two signatures, belonging to the persons in charge. It is the task of the employees at Invoice to check the presence of these two signatures, before entering the invoice into the system. Once entered, the invoice will be posted and paid afterwards by the system.

Whenever the business line that will carry the cost of the invoice works with SAP, the invoice will be entered immediately into the system. And in contrast to a PO-invoice, a nonPO-invoice will always be parked at a TPC, not only in case of lacking information. The corresponding TPC completes the invoice in terms of cost center, cost element and VAT rates because Invoice does not possess this information. When receiving back the invoice from the TPC, Invoice can post the invoice. This will create a work item at the inbox of the approving persons. This is according the same authorization tables as is the approval procedure at the stadium of creating a PO (see 3.3.2). It is only after the approving persons have signed electronically that the invoice will be paid.

In Figure 3.3 we find the invoice handling process visualized. Notice the different procedure for PO-invoices and nonPO-invoices.

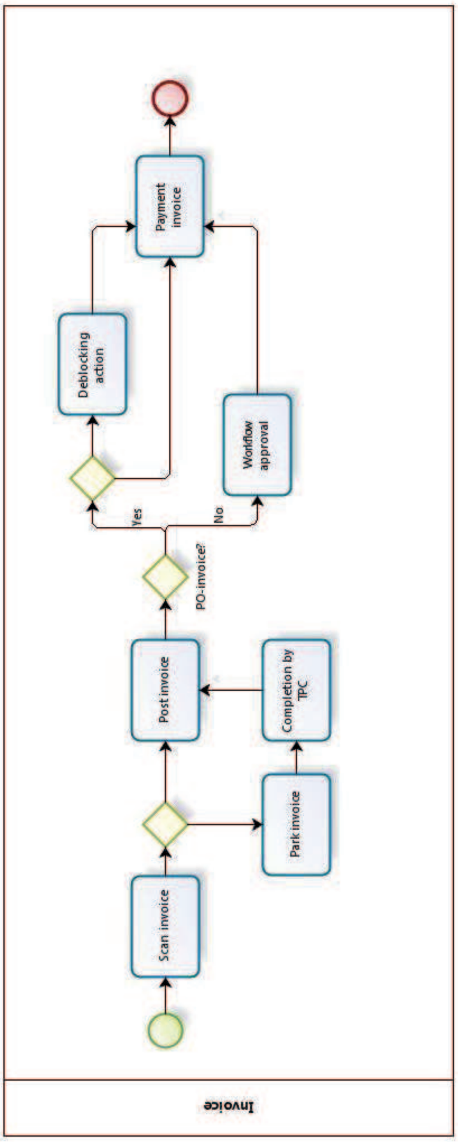


Figure 3.3: Invoice: handling of PO-invoices and nonPO-invoices

3.4 Risk Assessment at Epsilon

After conducting a process analysis, a risk assessment takes place to identify possible methods of committing fraud. This corresponds to the second component of COSO and is, in the light of risk, an important sequel of process analysis (Knechel, 2001). The risk assessment is approached the same way as the process analysis, this means by way of interviewing chief officers, observing employees executing their work within SAP, reading and questioning user guides, and discussing internal reports. Although the risks are assessed under close examination, the list may not be exhaustive. Also, the scope of this risk assessment is determined by the availability of data. For practical reasons, only the use of data of the ERP system is taken into consideration. In what follows, possible methods of practising fraud in the procurement cycle of Epsilon are exposed. The weak spots are grouped according to the department as where the fraud can be committed: Invoice and the Tactical Purchasing Cells. Three more possible frauds, not related to a specific department are explained separately.

3.4.1 Invoice

At Invoice, invoices enter Epsilon's system. As pointed out before, there are two main categories of invoices: PO-invoices and nonPO-invoices. Whenever it concerns a PO-invoice, approval has happened at the stadium of PO creation and each ordered item is only payable once. The ERP system prohibits double payment. Hence, fraud at Invoice level concerning PO-invoices is not considered a threat. NonPO-invoices on the other hand are more vulnerable to fraud. A nonPO-invoice only enters SAP at that moment, at Invoice. The approval still has to occur. This can be done in two ways: either the invoice will be approved electronically by two persons, or the invoice will be checked manually on two signatures by the employee at Invoice. In both cases personal enrichment is possible, mostly by collaborating with a supplier. But whatever the case is, the collaborating supplier must be present in the system. So if an employee wants to ensure personal enrichment, this condition has to be met and there are again two ways to do this: either one cooperates with an existing supplier of Epsilon, or one creates a phony supplier in the system. How this can be done, is treated next, as well as the possibilities of personal profit at Invoice, both handling nonPO-invoices electronically and manually.

Creating a Shell Company

Adding a supplier to the master record can only be done by a few authorized persons at Epsilon. These persons don't have authority to enter invoices or other related authorities. Separation of functions is implemented quite well at Epsilon. However, it is not impossible to make happen a new (phony) supplier is entered into the master record of suppliers. It is sufficient to create an invoice with the desired supplier coordinates and send it to Invoice. If there is nothing suspicious about the invoice (like the wrong address for example), the new supplier will be added by the authorized person.

Fraud Possibilities with Electronic Approval

To use ones occupation for personal enrichment at Invoice, handling nonPO-invoices that have to be approved electronically, one has a few possibilities. All depends on whether the employee works with an existing supplier, or a fictive supplier.

Cooperation with an existing supplier

If the employee is cooperating with an existing supplier, invoices can be entered twice, changing a little detail the second time. This change is necessary to circumvent the internal control, implemented to prevent mistakes. It is enough to add an extra character to the invoice number for example. This will not be noticed easily when comparing the original invoice with the information which has been entered into the system. When the supplier is a regular supplier, the persons that have to approve the purchase will not hesitate to approve an invoice more or less. They trust on their subordinates.

Another possibility is to send phony invoices. Instead of entering twice the same (legitimate) invoice, the supplier can send invoices for goods or services that were never delivered. This may be more difficult to track, since there will be no similar invoices that may look suspicious to investigators. The data now has to be compared with the reality of receiving goods etcetera. But typically, nonPO-invoices do not have a formal Goods Receipt, which makes fraud easier. On the other hand, once detected, proof will be more easy to gather, since there is hard evidence present (invoices for undelivered goods or services). The supplier takes a larger risk working this way than when an employee enters his invoices twice. In the latter case, the supplier can always pretend to be the subject of a simple mistake.

Using a shell company as supplier

After creating a shell company (see 3.4.1), invoices can be sent from that company to Epsilon. The approval of these invoices is harder to get. The approvers won't recognize the supplier and will ask questions about this new company or its goods. Of course, it could be possible that the first approver is in the fraud scheme as well. This makes it all a lot easier. Normally, if the first approver goes along, the second approver does so also, certainly if it's about small amounts of money.

Fraud Possibilities with Manual Approval

Personal enrichment with invoices where a manual approval is necessary, is much easier to obtain. The employee introduces invoices into the system, these don't even have to exist. One requirement for payment is again the presence of the supplier at the master record. Another requirement are two signatures on an approval form. However, this can not be checked by the ERP system itself.

Invoice Scanning

Invoice Scanning is a procedure, installed at Epsilon during 2005 and 2006. This procedure prohibits entering an invoice into the system, without actually receiving one, or entering it twice. All invoices received at Invoice are scanned, and the scanned documents are distributed electronically among the employees at Invoice. The information of the scanned document has to be typed into the right fields. This happens manually, and hence is still vulnerable to fraud. Yet, the possibilities of committing fraud with the Invoice Scanning procedure are limited. The dates of implementing the Invoice Scanning procedure have to be taken into account when analyzing the data.

Possible Selection of Frauds at Invoice

Looking at the possibilities to commit fraud at Invoice, one can make a selection of frauds for analyzing further. For this selection, inspired by using data out of the ERP system, we only take the cooperation with existing suppliers in consideration. The creation of shell companies is not typical for frauds at Invoice and are mostly to be detected manually, not by data analysis. As described above, an employee at Invoice can commit fraud by entering twice an existing invoice and receiving kickbacks from the supplier, or he can enter phony invoices in between a series of legitimate invoices. For an investigation by data analysis, only double payments are taken into account. This is inspired by a) the difficulty of detecting phony invoices by analyzing data of the invoices only (no stock information[2]) and b) the fact that phony invoices can be sent as part of external fraud, and could have nothing to do with occupational fraud and abuse. The data analysis for reducing this selected internal fraud risk is an example of 'reporting techniques' as described in Section 2.2. All kinds of related data are extracted out of SAP, attributes and reports are created. Analyzing these results led to the discovery of some double payments within Epsilon. However, reporting techniques are beyond the scope of this dissertation and therefor this study is not reported here.

3.4.2 Tactical Purchasing Cell

Invoice is not the only department where procurement fraud has a chance to occur. At the Tactical Purchasing Cells (TPC's), employees of Epsilon create Purchasing Orders and change them if necessary. In this process, there is a so called 'hot spot'.

An employee at a TPC is authorized to enter new purchasing orders, which have to be approved by two other persons. Only when approved, the PO can be sent to the supplier. There is no freedom here for the employee to do something irregular. However, the employee at a TPC has a way to (ab)use its occupation.

An employee at a TPC has the authorization to change a PO, even after the PO is approved and released. Mostly, this triggers a new work flow to approve this change. Small changes however, don't trigger this work flow. Small changes are for example altering the delivery address, but also altering the order itself. If this results in a small percentage change of the ordered value, the system does not require a new approval. This percentage depends on the total value. If the order is about 12,500 euros or less, an order change up to 5% is permitted to be made by an employee at a TPC without approval. If the order has a total value between 12,500 and 125,000 euros, there is only a 2% freedom in modification. An order above 125,000 euros can't be modified without approval.

It is clear that if an employee of Epsilon communicates these internal procedures

[2] Stock information is not captured in the ERP system at Epsilon

to a supplier, the two can set up a cooperation with the employee raising systematically each order after approval, and the supplier sending kickbacks to the employee afterwards. This hot spot will be taken into account when analyzing data.

3.4.3 2% margin at item level

A third fraud possibility is again the use of knowledge of internal procedures. Knowledge is power and power can be converted into money. This specific fraud can be committed by employees with different occupations within Epsilon. The knowledge we are talking about is the following: with PO-invoices, payment is blocked or not by matching the billed amounts on the invoice to the price at the PO times the quantities at the Goods Receipt document. If there is a match, the invoice will be paid automatically. As with other procedures, there is a margin on this rule. If there is a slight deviation on the price that was agreed upon times the quantity delivered, the invoice will still be paid. This deviation can go to 2% and still pass through the automatic payment system. An important remark on this margin has to be made however. This 2% rule is only valid per line item. So if the invoice is for eight different items (for example pencils, paper etc.), the 2% raise on the agreed price, can only pass the system if it is spread over the items. So an increase larger than 2% on one item and no increase on the remaining seven items, won't be paid.

This internal procedure again could be misused by sharing this information with a supplier and set up a cooperation. The employees benefiting from this abuse could be working on several departments. They theoretically don't even have to be part of the procurement cycle. Practically however, persons creating advantage of this knowledge would probably be involved in purchasing.

3.4.4 Service Desk

If an infrastructure problem arises, Epsilon employees can call to the service desk to get things repaired quickly. The service desk is authorized to order the specified services up to 600 euros, without following the normal approval procedure. This advantage could be exploited out of laziness, but it could also be used to mask orders for undelivered services. The delivery of services is rather difficult to check upon (it is often at another location) and hence a suited object to fraud. Part of the reward for the (undelivered) service could get back to the employee at the service desk. This is another fraud worth investigating. However, it is not suited in the light of data analysis, because it requires on the field investigation.

3.4.5 Collaboration of Employees

As already mentioned in previous scenarios, there are several fraud schemes relying on the collaboration of employees. These schemes can build on bilateral collaborations between employees or collaborations between employees and external parties. Examples of the latter are already presented in previous subsections. Examples of the former are combinations of employees working at a TPC and approvers. Because in both cases an internal party needs to be involved, this is classified as internal fraud, although an external party could be involved on top of this. However, these types of internal fraud are far more difficult to trace down. Even the suggested use of descriptive data mining will not redress this issue. This is because the result, a PO and a corresponding paid invoice, is the same as any other, legitimate PO and invoice. Also the behavior of the PO (times it is approved etc.) is not *per se* different than legitimate PO's. A solution may be found in investigating the business process itself, instead of the results of the process, but this is not under discussion now.

Chapter 4

Data Description

In Chapter 3 the purchasing process is described. In this chapter we take a closer look at the data this process creates, also the data used in this dissertation. Epsilon works with SAP and every transaction made in this environment creates information that is saved in sundry databases. First the purchasing process is repeated shortly with explanation of all the information saved during this process. This will be explained in Section 4.1. This information is collected during the process analysis, but it is for reasons of clearness that this information is bundled and described in this separate chapter. In the section following the Saved Data, something more is told about where these data are located in the system. The data is covered in a few tables and the structure of these tables is highlighted in Section 4.2.

4.1 Saved Data during the Purchasing Process

4.1.1 Creation of Purchase Order

The creation of a Purchase Order (PO) in SAP is the first trace a purchase can leave in the system (at Epsilon). "Can", because not all purchases evolve in this manner, as mentioned before. However, when a purchase is started in this way, this is the first moment the purchase enters the system. The employee working at a TPC, creating the PO, enters a lot of information into the SAP environment. All this information is logically saved somewhere. The information captured is two-dimensional.

The first dimension of information covers data entered in the header of the PO.

This is capital information concerning the PO and is captured in one table, EKKO. Fields like the PO number, type of PO, date of creation etc. are available in this table. After the header, the employee enters the items the requester wants to order. This is the second dimension of information and is captured in a separate table, EKPO. Each item covers one line with details about this item, like material group, order unit, order quantity, price per unit etc.

4.1.2 Release Strategy

After a PO is entered, a release strategy has to be followed to release this order. Each purchase is categorized in two ways: first the amount of the PO is critical for the release, secondly the business line that purchases. There are numerous different release strategies, depending on the purchasing group. Each strategy has its own boundaries and related persons to approve. The fact that there is no consistency within these strategies and that there is no history saved, makes information about the release strategy unusable. The only information about a release, usable in our analyzes, is the person that executes the sign or release. This user ID is to be found in the change log. The change log also has the double dimensionality of header information and item information. The header of a change log captures who caused the change, being the releaser in case of an approval (table CDHDR), when this has taken place and what kind of change it was (release in this case). In the table with the item details, being CDPOS, one can see which fields exactly were changed, along with the old and new values.

4.1.3 Goods Receipt

If a PO is released, the supplier receives the order and eventually the delivery of the good(s) or service(s) will be made. As seen in Chapter 3, a Goods Receipt will be entered into the system. At this moment, Epsilon recognizes the cost of the purchase officially in its accounting system. The term 'actuals' is used. Before, when there was only a PO available, the purchase was seen as a 'commitment'. Whenever actuals are posted, a financial document is created in the SAP environment. The information that such a financial document contains, is captured again in some tables. As with the PO's, this information is divided in the same two dimensions, each represented by one table, BKPF and BSEG.

The information of the header of a financial document is covered in the table BKPF. Fields like a unique document number (FD number), the date of the document and the type of the document are saved in this table. The type of the document tells us for example that we are dealing with a Goods Receipt. The second table, BSEG, encompasses information of the line items of the financial document. The lines of a financial document all together form the journal entry that is posted at the financial accounting system. The line items with regard to the actuals are linked to the line items of the PO. This kind of information is all captured in the table BSEG.

The fact that a Goods Receipt is entered into the system however, is captured in another table. This table encompasses the purchasing order history and tells us for each PO line whether a Goods Receipt is posted or not. This table in SAP calls EKBE.

4.1.4 Invoice Handling

After the goods are received, the invoice will follow. Invoices are handled at Invoice and are, like Goods Receipts, seen as financial documents. Based on this document, actuals will be posted and information is captured in the BKPF and BSEG tables. Whenever the purchase is preceded by a PO, the invoice is used to correct the actuals posted at the time of the Goods Receipt. For example: 10 units are ordered by a PO and a price of 34 each is agreed upon. At the time of creating the PO, a commitment of 340 is made. (information in EKKO and EKPO) If later on the Goods Receipt follows and indicates that 9 out of the 10 ordered units are actually delivered, SAP makes an estimate of the future actuals. This estimate is based on the quantity of the Goods Receipt (9) and the price agreed upon in the PO (34 each). As a result actuals of 9 times 34 are posted. At last the invoice will arrive for indeed only 9 units, but at a price of 34.1 each. At that moment, the actuals of 9 times 34, posted before, are supplemented with the additional 9 times .1. The information of this transaction is captured in the same way as the information of the Goods Receipt, whereas they are both financial documents. The document type will make clear which kind of document triggered the actuals, a Goods Receipt or an invoice.

An invoice can also be received without being preceded by a PO and a Goods Receipt. In that case, the total sum of actuals will be posted at the time of the receipt of invoice. The document type of such an invoice (a nonPO-invoice) will be different than of a PO-invoice. This information, amongst other information, is captured in

BKPF.

4.2 Structure of the Database

The seven most important tables with relevant information are already mentioned. Some linking has been explained too, but not all of it. In this section the structure of how the data in all tables are joined, will be explained.

4.2.1 Purchasing data at our disposal

Before setting out the underlying relations between the tables, the selection of data at our disposal is made clear. In Table 4.1, all data concerning purchasing at our disposal are listed. The location of the data is given by mentioning in which table the information is captured. Fields that are also available in other tables and will as a consequent be used as linking keys, are printed in italics. Notice that the five tables in SAP contain much more information. The listed fields (being the data dump at our possession) are only a selection out of the 730 fields available.

Apart from the tables mentioned in Table 4.1, there are CDHDR and CDPOS, containing header and item level information of changes. The available data of these tables is listed in Table 4.2. As can be seen, each documented change has a unique change number. This is the link between CDHDR and CDPOS. The object ID in CDHDR tells us what the subject of the change is. This can be any kind of document. Regarding to changes on a PO, the PO number will function as object ID. Regarding to changes on an invoice, the financial document number will be used. The saved date and employee ID speak for itself. The transaction code on the other hand reveals information about the kind of change that has taken place. This code can for example indicate whether the PO subjected to the change has been created, modified or approved. At the CDHDR table we can however not see which line items of the PO or invoice concerned are altered, nor can we see what the exact content of the document was on the date of the change. These are important remarks to consider before analyzing this data.

If we look at the CDPOS table, we get more information of each change (link through change number). We can find out which fields exactly are modified (and in which table these fields are located). We also see the old content and the new content of those changed fields. Using the table key we can find out which item line of the invoice or PO is changed. This link is not straightforward, because you have to look up the creation of the table key for each kind of object ID in CDHDR. For example, when the object ID in CDHDR shows the change concerns a PO, the table key in CDPOS is a combination of the company code, the PO number and the item line. When the object ID in CDHDR refers to an invoice, the table key at CDPOS may for example be a compression of company code, client number, invoice number, item line and fiscal year. However, after looking into the table key, it is possible to link the item information of changes to a PO or invoice item line.

Table 4.1: Purchasing data available - selection

—BKPF—	BSEG—	EKKO —	EKPO —	EKBE—
FD number	*FD number*	*PO number*	*PO number*	*PO number*
document type	FD line number	PO type	*PO line number*	*PO line number*
document date	posting key	status	last changed on	document number
posting date	account type	date of creation	deletion indicator	item number doc.
entry date	D/C indicator	document date	order quantity	category
time of entry	amount	employee ID	order unit	posting date
changed on	assignment nr	last item	net value	quantity
employee ID	cost center	vendor nr	gross value	amount
status	general ledger nr	payment terms	print price indicator	D/C indicator
unplanned delivery cost	vendor nr	purchasing group	estimated price indicator	delivery completed
object key	discount	vendor name	GR indicator	reference
reverse doc nr	quantity	release indicator	GR non valuated ind.	ref. doc.
	order unit		invoice receipt indicator	entry date
	PO number		GR-based invoice	entry time

Continued on next page

—BKPF—	BSEG—	EKKO —	EKPO —	EKBE—
	PO line number clearing date		outline agreement nr item nr outline agreement	user id

Table 4.2: Data available in CDHDR and CDPOS

CDHDR	CDPOS
object ID	*change nr*
change nr	change item
date	table name
employee ID	table key
transaction code	field name
	old value
	new value

Now the data available are known, the relations between the tables can be made. This will be done in the next section.

4.2.2 Relations between the data tables

Getting some insight in the relations between the tables is fundamental in understanding the data structure. The relations are explained below and visualized in Figure 4.1.

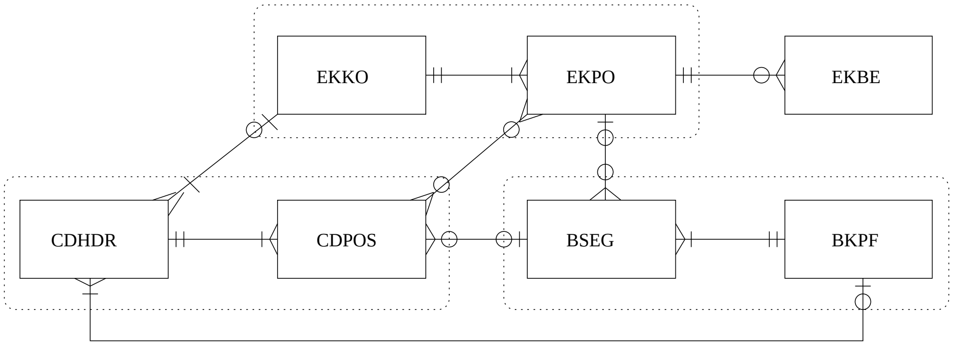


Figure 4.1: Relation between tables

For every record in BKPF (containing financial documents headers), there exists at least one and at most n item lines in BSEG. On the other hand, a record in BSEG belongs to one and only one record in the header table BKPF. This is obvious since a financial document must exist at least out of one line and every line belongs to one and only one financial document.

A record of BSEG can correspond to maximum one, minimum none records of the PO item table EKPO. If there is no match with an item line of EKPO, we are dealing with a nonPO-invoice. If there is a match, this is a PO-invoice. In the other direction however, a record of EKPO can match to maximum n (not one!), minimum none records of BSEG. The match with more than one record of BSEG will be the case when there is more than one financial document linked to the PO. These documents can be for example a goods receipt and an invoice. If there is however no match with a record of BSEG, there is no financial document yet for this PO.

For every record of EKPO there exists one and only one match in EKKO. For every record in EKKO, there can be more than one but at least one match in EKPO. These relations are obvious since a PO can exist out of one or more item lines and each item line belongs to one and only one PO. Also, for each line in EKPO, there can be minimum no and maximum n corresponding lines in EKBE. Here we can see that the Goods Receipts and the invoices are linked to an item line of a PO and not to a PO as a whole.

Concerning the changes, we can link both EKKO and BKPF to CDHDR, and going from there to CDPOS. CDPOS itself is linked with the item level tables BSEG and EKPO. On each item line of a PO or invoice, can be none or n changes performed. In the other direction is a line item of a change linked to exactly one line item of another document. This other document can be a PO or an invoice, or some other document, not included in our data set. On header level, the change also concerns always one document, whether it is a PO, an invoice or some other document. Because a document can only exist after a creation, which is seen as a change in the change log, for each document there is at least one change and maximum n changes. All these relationships are depicted in Figure 4.1. Because traditional data mining will be applied -as opposed to multi-relational data mining- a flat file has to be made, using the known relations underlying the relational database.

Chapter 5

Descriptive Data Mining Application

In this chapter an application of the IFR^2 Framework is presented. For this application, we had the corporation of Epsilon, as mentioned in Chapter 3. As a refresh, the framework is presented another time in Figure 5.1. The following sections are the result of implementing this framework at Epsilon.

5.1 Data Set Description

The first step in the IFR^2 Framework is to select a business process with advanced IT integration. As expounded in Chapter 3, the procurement business process at Epsilon was selected for internal fraud risk reduction. This section provides more insights in the data available and presents several descriptives.

Data transactions from Epsilon's procurement cycle were collected. Epsilon provided a txt-dump from their SAP system. All PO's that in 2006 resulted in an invoice are the subject of our investigation, yielding a data set of 36.595 observations. This raw data is then reorganized into appropriate tables to support meaningful analysis. After the creation of these new formats, additional attributes were created as enrichment. Based on domain knowledge and supported by descriptive statistics, a pre-clustering step is made. PO's are split in two groups: old PO's and new PO's. Old PO's are the ones created before July 2005. The fact that they are included in our data is because an invoice of the year 2006 can be linked to a PO created in

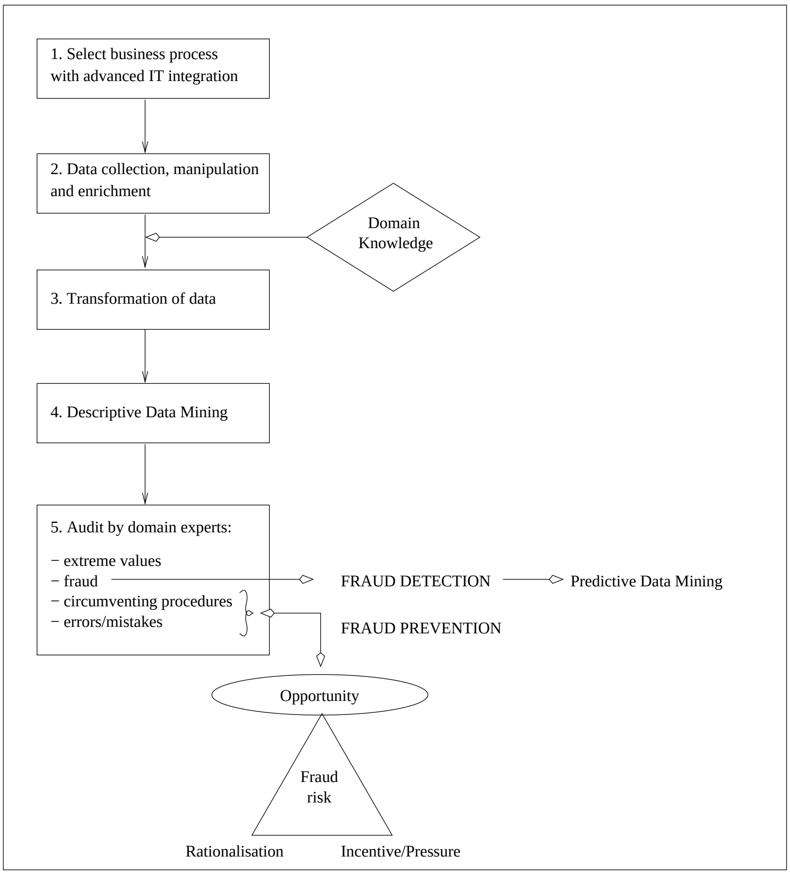

1. Select business process
with advanced IT integration
2. Data collection, manipulation
and enrichment
Domain
Knowledge
3. Transformation of data
4. Descriptive Data Mining
5. Audit by domain experts:
− extreme values
− fraud
− circumventing procedures
− errors/mistakes
FRAUD DETECTION
Predictive Data Mining
FRAUD PREVENTION
Opportunity
Fraud
risk
Rationalisation
Incentive/Pressure

Figure 5.1: The IFR^2 Framework

2005 or even before 2005. However, if a PO is from before July 2005 (keep in mind that even still in 2006 invoices were linked to this PO), this PO shows a different life cycle than if it were younger. While a 'new' PO will more probably have a life cycle of creation, approval and an attached invoice, an 'old' PO will probably be modified more often in between and have several invoices attached to it. The subset of old PO's contains 2.781 observations while the subset of new PO's counts 33.814 observations. Both subsets of PO's were subjected to the main part of the methodology (step one through four). Because the latter group is the most prominent in assessing internal fraud risk (most recent and highest value in terms of fraud prevention) and given its magnitude, this chapter gives only detailed test results of the new PO's. The other side of the picture is that this large data set poses more problems in the fifth step of our methodology, namely the auditing of interesting observations. We restrict this study to provide recommendations on this matter for the new PO's. For the subset of old PO's however, the audit step is effectively executed and these results will be reported after the discussion of the new PO's. In what follows, we provide descriptive statistics, univariate and multivariate analysis results and recommendations on the fifth step for the subset of new PO's. Although all these steps are also executed on the subset of old PO's, we restrict to the reflection of only the fifth step for this subset. In the following part, the term 'data set' refers to the subset of new PO's unless stated otherwise.

The most important attributes to describe a PO and its life cycle are the following: the name of the creator, the supplier, the purchasing group, the type of purchasing document, the number of changes, the number of changes after the last release and the number of price related changes after the last release. 'Changes' are 'events' stored in the log file of the ERP system, so it should not be mistaken for changes in the sense of modifications alone. The creation of a PO, the modifications of a PO, the signs (first approval) and the releases (second approval) are all logged as 'changes'. Concerning the categorical attributes, there are 91 creators in the data set, 3.708 suppliers, 13 purchasing groups and 6 document types. (see Table 5.1) The histograms of Figure 5.2, 5.3, 5.4 and 5.5 provide us with some insights of the distribution on these attributes.

As can be seen in Figure 5.2 the 91 creators do not introduce the same number of PO's in the ERP system. This is caused by the individual characteristics of each purchase. Some creators are responsible for a particular type of purchase which results in entering a lot of PO's, while other creators are responsible for other types of

Table 5.1: Categorical attributes.

Categorical	Recurrence in data set
Creator	91
Supplier	3.708
Purchasing Group	13
Document Type	6

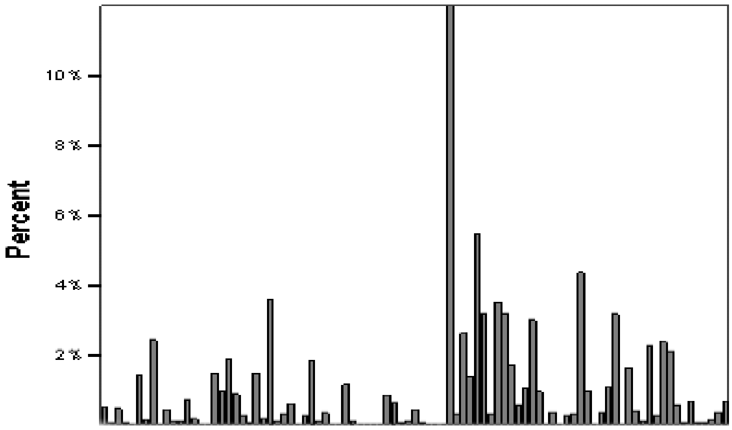


Figure 5.2: Distribution of creators in data set.

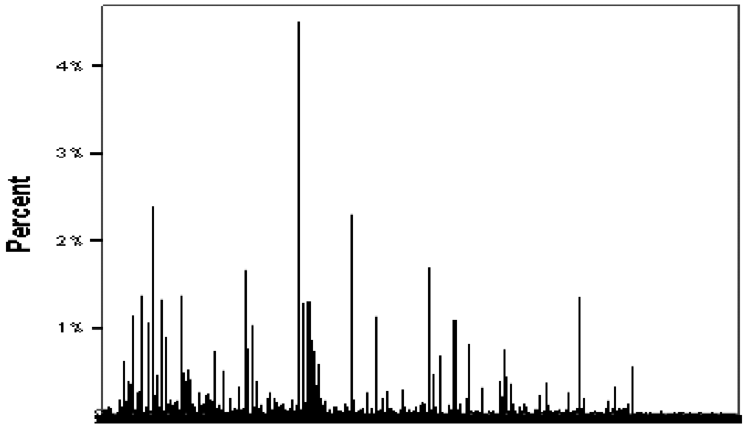


Figure 5.3: Distribution of suppliers in data set.

purchase which involves processing only a few PO's. There is one 'creator' responsible for 25% of the PO's in the data set.[1] This is however not a person, but concerns a

[1] This is beyond the scale of this graph, in order to remain the visual variation in distribution

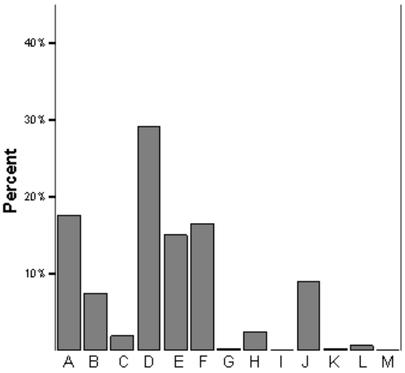


Figure 5.4: Distribution of purchasing groups in data set.

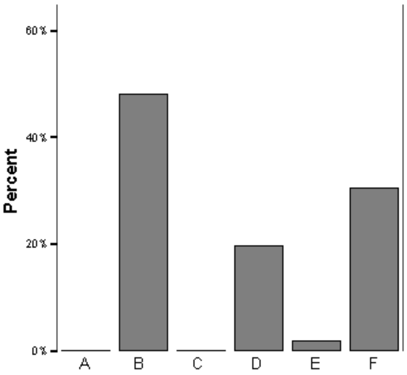


Figure 5.5: Distribution of document types in data set.

method of creating a PO, namely by inputting a batch into the ERP system. SAP sees this method as one creator. Also the turnover in terms of personnel has its reflection on the number of PO's per employee. For example an employee that is hired (or retired) halfway 2006, will process fewer PO's than employees registered for a full year.

Like creators, the frequency of suppliers in the data set is liable to the specific characteristics of the product or service supplied. There are for example more PO's concerning monthly leasing contracts for cars than there are for supplying desks. Hence the former supplier will be more frequently present in the data set than the latter. Concerning the 13 purchasing groups, there is no difference in expected fraud risk between the different groups. Some groups are more present than others in the data set, but this can all be explained by domain knowledge. The same goes for the six different purchasing document types. The run-down of two types (A and C) is clearly visible. The remaining four types have their specific characteristics, but there is no expected difference concerning fraud risk.

The numerical attributes are described in Table 5.2. For each attribute, three intervals were created, based on their mean and standard deviation. For the first attribute, the intervals were [2-4], [5-8] and [9-...], for the second attribute [0-0], [1-2] and [3-...] and for the last attribute [0-0], [1-1] and [2-...]. In Table 5.2 we see that

among the other creators.

there is a highly skewed distribution for the three attributes, which is to be expected for variables that count these types of changes. The changes are supposed to be small in numbers.

Table 5.2: Descriptives of numerical attributes.

Attribute	Minimum	Maximum	Mean	Standard deviation	1st interval frequency (%)	2nd interval frequency (%)	3rd interval frequency (%)
Number of changes	1	152	4.37	3.846	71.3	21.5	7.2
Number of changes after last release	0	91	.37	1.343	80.9	11.9	7.2
Price related number of changes after last release	0	46	.15	.882	91.1	6.7	2.2

5.2 Latent Class Clustering Algorithm

For a descriptive data mining approach, we have chosen a latent class (LC) clustering algorithm. We prefer LC clustering to the more traditional K-means clustering for several reasons. The most important reason is that this algorithm allows for overlapping clusters. An observation is provided a probability to belong to each cluster, for example .80 for cluster 1, .20 for cluster 2 and .00 for cluster 3. This gives us the extra opportunity to look at outliers where the observation does not really belong to any cluster at all. This is for example the case with probabilities of: .35, .35 and .30. Another reason is that LC clustering algorithm has the ability to handle attributes of mixed scale types and has information criteria statistics to determine the number of clusters. For a more detailed comparison of LC clustering with K-means, see Magidson and Vermunt (2002).

In LC analysis, one starts with the idea that any dependency between the observed or manifest variables can be explained away by some other variable(s). These other variables can be unobserved or unobservable, called latent. We believe internal fraud risk can be represented by such a latent variable and hence fraud risk can be deduced from available information, i.e. manifest variables. We have technical information about a PO (who made it, when, ...) and we have operational information about this PO (how many times is it changed, ...). This operational information describes a behavior. It is this behavior, in combination with technical information, that we believe leads us to fraud. The main objective is to move from technical data over behavioral data (third step) to behavior (fourth step). Particular this challenge stimulates the use of data mining techniques, in that data mining enables us to recognize patterns we are unaware of. These patterns are used to gain insights in the behavior of people.

In LC clustering, a specific type of LC analysis, objects are assumed to belong to one of a set of K latent classes, with K being unknown. Observations in the same class are similar in the probability distributions underneath the manifest variables' scores. It is assumed that a population is a mixture of underlying probability distributions. Parting these different distributions provides us different clusters. The latent variable(s) are believed being capable of doing this.

The basic model for LC clustering has the form

$$f(\mathbf{y}_i|\theta) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i|\theta_k)$$

$\mathbf{y}_i$ denotes an observation i's scores on the dependent variables. π_k denotes the prior probability of belonging to latent class (or cluster) k. This model puts the conditional distribution of $\mathbf{y}_i$ (given the model parameters of θ) as a mixture of class-specific densities, $f_k(\mathbf{y}_i|\theta_k)$.

Since there is the assumption of local independence between the manifest variables, this can be rewritten as

$$f(\mathbf{y}_i|\theta) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} f_k(y_{ij}|\theta_{jk})$$

where J denotes the total number of dependent variables and j is a particular dependent variable. Instead of specifying the joint distribution of $\mathbf{y}_i$ given class membership, this is split up into separate univariate distribution functions for each y_j.

The latent variables are assumed to explain all associations between manifest variables (so that there is local independence between them). The goal is to determine the smallest number of latent classes K that is sufficient for safeguarding this assumption.

The standard procedure is to test Model H_0, with $K = 1$, first. From this model on, latent classes are added.

For assessing the fit of LC models, there are several criteria available. The most widely used approach is the use of the statistic L^2. The lower this L^2, the less probability the model fits the data by chance. If L^2 equals 0, this means the variables are perfectly independent from each other and all associations among the manifest variables are explained by the latent variables. For selecting a model, information criteria are also quite popular. The ones most used are the Akaike, the Bayesian and the consistent Akaike information criteria, or AIC, BIC and CAIC. These criteria are based upon the log-likelihood (LL). Where LL always gets better (i.e. closer to 0) when K is raised, the information criteria take the number of parameters (and BIC also the number of degrees of freedom (df)) into account. The general definitions of BIC, AIC and CAIC are given below:

$$\begin{aligned} BIC &= -2LL + ln(N)M & (5.1) \\ AIC &= -2LL + 2M & (5.2) \\ CAIC &= -2LL + [ln(N) + 1]M & (5.3) \end{aligned}$$

with N being the sample size and M the number of parameters. The smaller the criterion, the better the model.

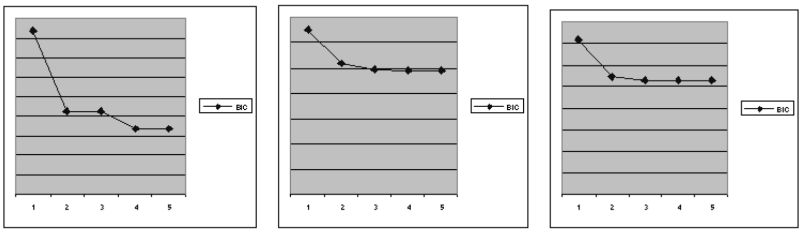
The mode of operation of starting with Model H_0 and building further, is inspired by comparing the information criteria values with these of Model H_0.

For more and detailed information about LC analysis, we refer to Kaplan (2004) and Hagenaars and McCutcheon (2002).

5.3 Univariate Clustering

5.3.1 Model Specifications

Before turning to the core of our model of applying a descriptive data mining approach on behavior describing attributes, we apply univariate clustering to explore data. The univariate analysis is applied on obvious attributes. The three most obvious attributes were selected: number of changes (Model A), number of changes after release (Model B) and number of price related changes after release (Model C). For each attribute and its belonging univariate clustering model, we executed the LC clustering algorithm with the number of clusters (K) set equal to 1 till 5. This yielded the BIC information criteria values plotted in Figure 5.6. The AIC and CAIC values showed the same pattern. As you can see, the BIC values drop three times heavily until the 2-cluster model. Beyond the 2-cluster model, the decreases are more modest. Only at Model A, the 4-cluster model is an alternative candidate. The classification statistics of this model are however less satisfactory than those of the 2-cluster model. Based on these values, we decide to use three times the 2-cluster model for further exploration.



(a) Model A (b) Model B (c) Model C

Figure 5.6: BIC values univariate clustering.

5.3.2 Results

In Table 5.3 we find the clustering results for Model A, B and C with the cluster size in percentage and the mean value of the clustering attribute for each cluster. For each of the three univariate models, one large and one small cluster is given as output of the 2-cluster models. The small clusters contain 1,428, 1,085 and 344 PO's (out of the 33.814) respectively for Model A, B and C. The 344 cases of the small cluster of Model C are fully incorporated in the 1,085 PO's of Model B. This is the only classification consistency between the three models. Except for the cluster size, the mean value of the attribute in each cluster is given. For each model, cases with a small value (of the count attribute) are classified in the large cluster while the small cluster is characterized by a higher mean value of that attribute.

The clustering in each model is based on the value of one attribute. By looking

Table 5.3: Results of univariate clustering.

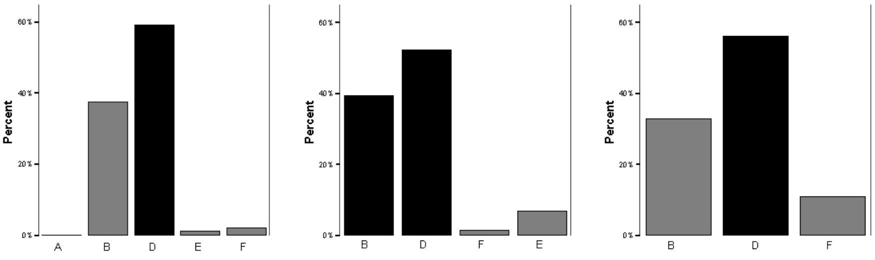
Model A		
Number of changes		
	Cluster 1	Cluster 2
Cluster size (%)	0.96	0.04
Mean	3.84	16.79
Model B		
Number of changes after last release		
	Cluster 1	Cluster 2
Cluster size (%)	0.95	0.05
Mean	0.17	3.82
Model C		
Price related number of changes after last release		
	Cluster 1	Cluster 2
Cluster size (%)	0.99	0.01
Mean	0.09	5.01

at the following four attributes, we get insight in which kind of PO's are classified in these small clusters, aside from the high score on the clustering attribute: document type, purchasing group, creator and supplier.

In Figure 5.7 we see the distribution of the document types in the small clusters of Model A, B and C. Model A and C both highlight document type D, while at Model B both B and D hold a prominent place.

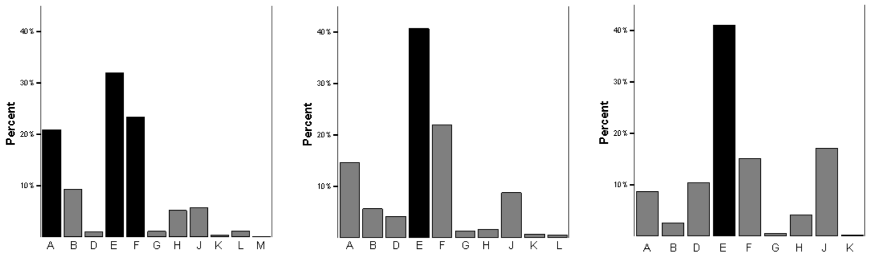
In Figure 5.8 we see the distribution of the purchasing groups in the small clusters of Model A, B and C. Model A highlights three purchasing groups (A, E and F), while Model B and C only put E in the spotlights.

Looking at outliers concerning creators and suppliers, we can see that there is again some overlap, but no complete consistency among the three models. In Figure 5.9 we find a top three of creators, named C1, C2 and C3. Building further on Model



(a) Model A (b) Model B (c) Model C

Figure 5.7: Distribution of document types in small univariate clusters.



(a) Model A (b) Model B (c) Model C

Figure 5.8: Distribution of purchasing groups in small univariate clusters.

A, these are the creators interesting to have a closer look at. If we use another attribute to cluster on, for example the clustering attribute of Model B, we get another composition of creators in the small cluster. This distribution is presented in Figure 5.10. One creator out of Model A's top three returns with a frequency of 13.4%, namely C2. (This is beyond the scale of this graph, which is set equal to other graphs concerning the frequency of creators.) In spite of this consistency between Model A and B, Model B highlights another top three. The two new creators are named C4 and C5. Turning to Model C, C2 is again represented and now in an even bigger percentage of 25.6%. Also C4 of Model B returns, but again three new creators stand in the spotlight, C6, C7 and C8. So depending on which attribute we choose to perform the univariate clustering, other creators get our attention.

The same analysis is made of the suppliers. The small cluster of Model A highlights

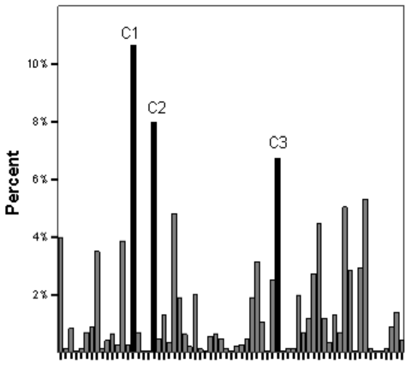


Figure 5.9: Distribution of creators in small cluster of Model A.

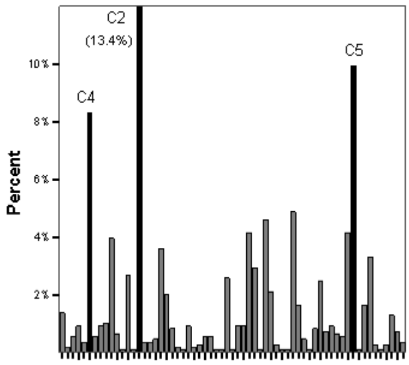


Figure 5.10: Distribution of creators in small cluster of Model B.

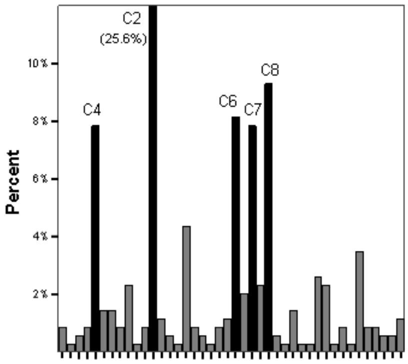


Figure 5.11: Distribution of creators in small cluster of Model C.

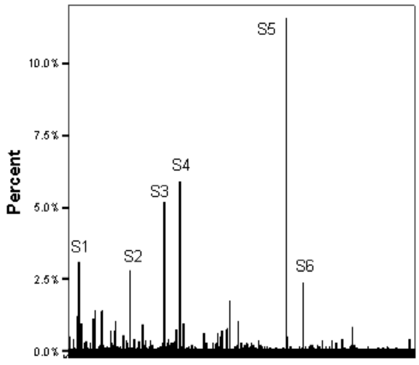


Figure 5.12: Distribution of suppliers in small cluster of Model A.

a top six, S1 till S6. In the small cluster of Model B, two of those six suppliers, S5 and S6, are again represented in the top three along with the new supplier S7. In yet another composition of the small cluster, based on the clustering attribute of Model C, two more new suppliers get attention, and some other don't. A top three arises, with S8 and S9 being new suppliers, together with S3 from the top six out of Model A.

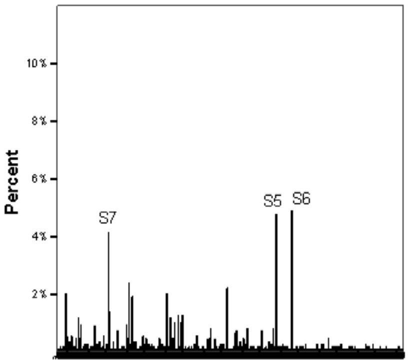


Figure 5.13: Distribution of suppliers in small cluster of Model B.

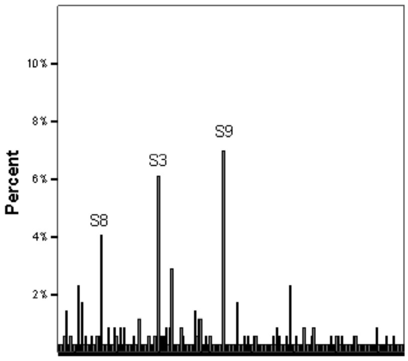


Figure 5.14: Distribution of suppliers in small cluster of Model C.

5.4 Multivariate Clustering

5.4.1 Model Specifications

The univariate clustering yielded contradictory information, depending on which attribute was taken to cluster on. A multivariate analysis takes several attributes at the same time into account. Before we can apply this analysis, we have to execute the third step of our methodology, namely to translate technical data into attributes that describe behavior. For performing this step, we take into account the particular type of fraud risk we wish to reduce. The fraud risk linked with entering PO's into the ERP system is connected with the number of changes one makes to this PO, and more specifically, the changes made after the last release. There is namely a built-in flexibility in the ERP system to modify released PO's without triggering a new release procedure. For assessing the related risk, we selected four attributes to mine the data. A first attribute is the number of changes a PO is subjected to in total. A second attribute presents the number of changes that is executed on a PO after it was released for the last time. A third attribute created is the percentage of this last count that is price related. So what percentage of changes made after the last release is related to price issues? This is our third attribute. A last attribute concerns the magnitude of these price changes. For expressing this magnitude, we calculate the mean of all price changes per PO and its standard deviation. On itself, no added value was believed to be in it. Every purchaser has its own field of purchases, so cross sectional analysis is not really an option. However, we combine the mean (μ) and standard deviation (σ) to create a theoretical upper limit per PO of $\mu + 2\sigma$. Next, we count for each PO how often this theoretical limit was exceeded. This fourth attribute is taken into account in our data mining approach. In this core model, no categorical attributes were added. As a robustness check however, attributes like document type and purchasing group were included in the model. The results did not significantly change by these inclusions.

After the selection of attributes, we need information to set the value of K. We therefor execute the LC clustering algorithm with K set equal to 1 till 5. This yields the BIC values plotted in Figure 5.15. The BIC values drop heavily until the 3-cluster model. Beyond the 3-cluster model, the decreases are more modest. Based on these values, we decide to use this 3-cluster model.

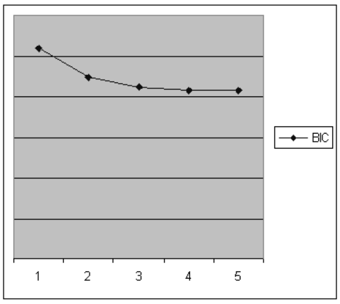


Figure 5.15: BIC values multivariate clustering.

5.4.2 Results

The profile of the 3-cluster model is presented in Table 5.4. It gives the mean value of each attribute in each cluster. To compare with the data set as a whole, the mean values of the population are also provided.

Looking at the profile of the 3-cluster model, there is an interesting cluster to notice, the third cluster, if it was even only for its size. Cluster 1 comprehends 76.6% of the total data set, cluster 2 22.1% and cluster 3 only 1.25%. Why is there 1.25% of all PO's behaving differently than the remaining PO's? Regarding the mean attribute values of this small cluster, this cluster is, besides from its size, also interesting in

Table 5.4: Profile of data set and 3-cluster model.

	Population	Cluster 1	Cluster 2	Cluster 3
Cluster size	100	0.7663	0.2212	0.0125
Number of changes	4.37	3.3378	6.7608	25.459
Changes after release	0.37	0.0193	1.2376	6.1257
Percentage price related	0.0756	0	0.3185	0.4094
Count over limit	0.01	0.0072	0.0194	0.2725

terms of fraud risk. The mean number of changes per PO in this cluster, is 25, as opposed to a mean number of changes of 4 in the data set. Why are these PO's modified so often? Not only are these PO's changed so much in their entire life cycle, they are also modified significantly more after they were last released (6 times) in comparison with the mean PO in the data set (0.37 times). These are odd characteristics. The mean percentage in cluster 3 of changes after the last release that is price related is also the highest percentage of the three clusters (40.9%). All together this means that the average PO in cluster 3 is changed 25 times in total, of which 6 changes occur after the last release and 2.4 of those 6 changes are price related. Concerning the magnitude of the price related changes, we can conclude that these changes of PO's in cluster 3 are more often much larger than the average price change in that PO if we compare this with price related changes of PO's in the other clusters. In cluster 3, there are on average 0.2725 price related changes larger than $\mu + 2\sigma$ per PO, in comparison with 0.0072 and 0.00194 per PO in cluster 1 and 2 and 0.01 changes in the entire data set.

Taking these numerical characteristics into account, one can conclude that cluster 3 has a profile with a higher fraud risk than the other two clusters.

Numerical attributes tell us that cluster 3 carries a fraud risky profile, but also categorical attributes behave in a different fashion than they behave in the data set as a whole. So there are the creators of the PO. One person for example created 39 out of the 408 PO's from cluster 3 (hereby representing 9.56% of cluster 3), while the same person only created 131 out of the 33.814 PO's, which counts only for 0.39% of the entire data set.

For calculating the probability of taking this person (called xxx) by chance 39 times of 408, given the prior distribution, we use the hypergeometric distribution. This looks as follows.

$$h_m = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$$

The hypergeometric distribution is a discrete probability distribution that describes the number of successes m in a sequence of n draws without replacement, given a finite population N with M successes. In our situation concerning person xxx this leads to:

$$h_{39} = \frac{\binom{131}{39}\binom{33.814-131}{408-39}}{\binom{33.814}{408}} < 1^{-15}$$

So if we select 408 cases at random out of the population of 33.814 observations, there is a probability less than 1^{-15} that we pick 39 cases with user-id xxx, given the prior distribution of 131 successes in the population. This event is very unlikely to happen by coincidence.

Not only creators made such significant increases in representation, but also some suppliers are significantly more represented in cluster 3 than they are in the full data set. We screened all creators and suppliers on significant increases in representation between the data set and cluster 3 with a significance level of $h < 1^{-5}$. 14 suppliers and 12 creators met this criterium. Not all of them are however equally important since an increase of 0.03% representation to 0.98% is not as impressive as an increase of 1.47% to 7.6%. Table 5.5 gives us more insights into the importance of the 14 suppliers and 12 creators.

Not only the creators and suppliers, but looking at Figure 5.16 and 5.17 we also find the distributions of purchasing groups and purchasing document types in cluster 3 differing from the distributions in the population. Purchasing group E stands out with its 39.2% in cluster 3, while it was purchasing group D that was highly represented in the total data set (29.1%). Concerning the document types, type B was found most prevalent in the total data set (48.1%), while type D was most prevalent in cluster 3 (64.0%).

To summarize these results, we find a small cluster with a high fraud risky profile, due to the large values for the behavior describing profiles. When looking at this small cluster in terms of creators, suppliers, purchasing groups and document type,

Table 5.5: Descriptives of creators and suppliers with a significant higher representation in cluster 3.

Representation (r) in cluster 3	Number of suppliers	Number of creators
$r < 1\%$	4	
$1\% < r < 2\%$	4	
$2.2\% < r < 4.5\%$	3	
$6\% < r < 7.5\%$	3	
Total	14	
$r < 2\%$		3
$2.9\% < r < 3.5\%$		3
$5\% < r < 10\%$		6
Total		12

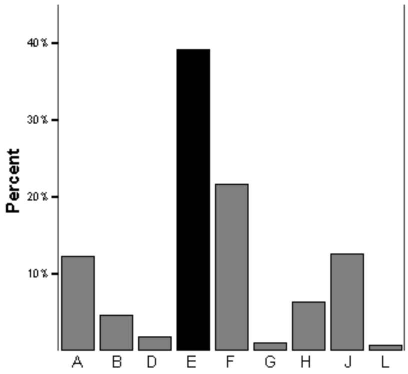


Figure 5.16: Distribution of purchasing groups in cluster 3.

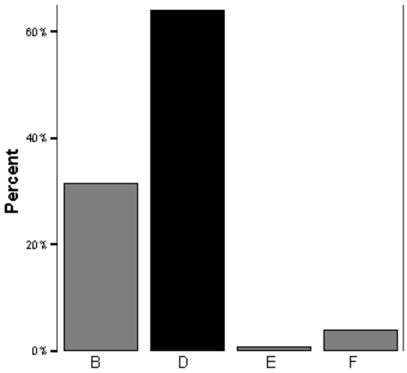


Figure 5.17: Distribution of document types in cluster 3.

we find other patterns than in the total data set. It would be interesting to have a closer look at these PO's to find answers on why they behave so differently from the total data set. This would be the fifth step of our methodology.

5.5 Audit by Domain Experts

Because it is too time consuming to audit all 408 PO's of cluster 3, it can be interesting to take a sample of PO's that are made by one of the creators described above or involve one of those suppliers (or both). In this context a smaller sample of cluster 3 was extracted by taking only those PO's of the six creators or in which one of the three suppliers that are most represented in the cluster were involved. This yielded a sample of 38 PO's. Why did they induced more PO's in this small cluster than in the other two clusters? What made these purchases this risky? Also the recurrence of a particular purchasing group (E) and purchasing document type (D) can shed an interesting light on deciding which PO's to audit. Auditing this kind of PO's can learn the company a lot about the opportunities that exist to commit fraud, in view of the fraud risky profile that the behavioral attributes describe. However, as already mentioned, the audit step is not executed for the subset of new PO's (cluster 3). At least it is not performed on the data of the small cluster. The possibility LC clustering provides to select observations that do not clearly belong to one cluster (see Section 5.2) is also explored. This however did not seem to be very interesting in this case. 42 PO's were identified as 'cases between clusters' and audited in-depth (the fifth step). The audit resulted in a few questions with regard to the use of the ERP system or procedures. Nothing however showed misuse of procedures.

The entire methodology, provided in Figure 5.1, is also applied on the subset of old PO's. The results of the descriptive data mining step are similar to the discussed results. The multivariate analysis also revealed a small fraud risky cluster. The small interesting cluster (in perspective of a fraud risky profile) of old PO's only contained 10 observations, with nine of them stemming from the same purchasing group and six of them created by the same employee. These 10 observations were audited in depth by employees of the investigations department. This examination occurred in close collaboration with the internal audit department and the department responsible for the SAP settings concerning Material Management. All changes that ever took place on the selected PO's are collected and looked into in detail. Also all accompanying invoices and, if applicable, corresponding e-mails were examined. The results of their investigation are summarized in Table 5.6.

Nine PO's, the ones in one particular purchasing group, are created and modified all over and over again. This is against procedures and makes investigating these PO's very difficult. By creating such complex histories of a PO, the opportunity of

Table 5.6: Summary of investigation by domain experts.

Category	Number of cases
Extreme values	0
Fraud	0
Circumventing procedures	9
Errors/Mistakes	1

committing fraud increases. Only insiders can unravel what really happened with these PO's, since they are such a mess. This of course increases the opportunity and risk of internal fraud. Also, the investigation of this practice has put things in another perspective concerning the separation of functionalities. A follow-up investigation by the audit and investigations department of the case company for this matter is approved. This follow-up will include interviews with the responsible chief officers and if necessary a reassessment of the procedures in order to adapt these toward the existing needs.

In the tenth PO a mistake is made. As explained before, a mistake that stays unnoticed creates a window of opportunity for internal fraud. The employee that first makes a mistake by accident, can afterwards consider how to turn this opportunity to one's advantage.

These are very good results in the light of internal fraud risk reduction, because all investigated PO's are contributing to fraud prevention by laying bare opportunities. Also, by investigating the 10 selected observations, additional odd practices came to light, which induced extra investigations. On top of this, the case company gave priority on auditing the procurement cycle in depth.

5.6 Multivariate versus Univariate Analysis

The results of using a multivariate descriptive data mining approach based on behavior describing attributes, provides us with interesting results. In the smaller subset of old PO's we encounter PO's that are changed over and over again. Also in the larger subset, changing the PO a lot of times is a primal characteristic of the selected observations. However, one could wonder if this outcome was not much easier to obtain, simply by applying univariate clustering instead of multivariate clustering. We do not go into the discussion about one method being generally better than another. What we can and want to say however, is that in our case, we did not find the same results by using a univariate LC clustering algorithm as we found by applying the multivariate LC clustering algorithm. Firstly, if we have a look at the profiles of all small clusters, both from the univariate as from the multivariate analysis (Table 5.7)[2], the profiles of the univariate small clusters are not as marked as the small cluster 3 we discussed. The marked profile is of high importance in order to make a narrow selection of cases for further auditing in the light of internal fraud risk. The small cluster of Model C is only for 43% incorporated in cluster 3, so this is not a selection of some core of cluster 3. We can conclude that the multivariate aspect of our analysis was indispensable to come to the presented profile.

Table 5.7: Profiles of small clusters from Model A, B, C and the multivariate model.

	Model A	Model B	Model C	Multivariate
Cluster size (PO's)	1,428	1,085	344	408
Number of changes	(16.87)	12.13	17.88	25.46
Changes after release	2.51	(5.18)	7.73	6.13
Percentage price related	0.27	0.40	0.84	0.41
Count over limit	0.09	0.06	0.17	0.27

When we look at the four categorical attributes purchasing group, document type, creator and supplier, the univariate samples do not present the same results either. Concerning the purchasing group, with E most prevalent in cluster 3, only Model B showed a comparable distribution of purchasing groups in its small cluster, although Model C also highlighted purchasing group E, but in another distribution profile. The same comparison can be made for the document type. This time Model A shows a comparable distribution of document types in its small cluster with the distribution in cluster 3, but Model B and C do not. Model B even put both document types B and D in the spotlights, instead of only document type D. Regarding creators and

[2] The values between brackets are of the univariate clustering attribute and should not be taken into account when assessing the resulting profile.

suppliers, the distributions are presented in Figure 5.18 and 5.19.

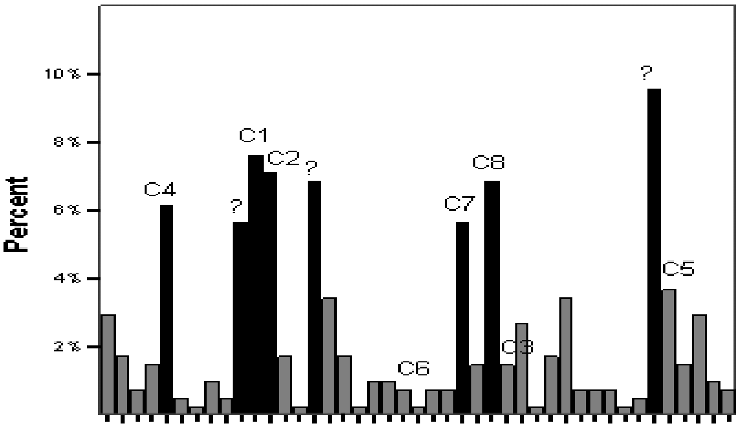


Figure 5.18: Distribution of creators in cluster 3.

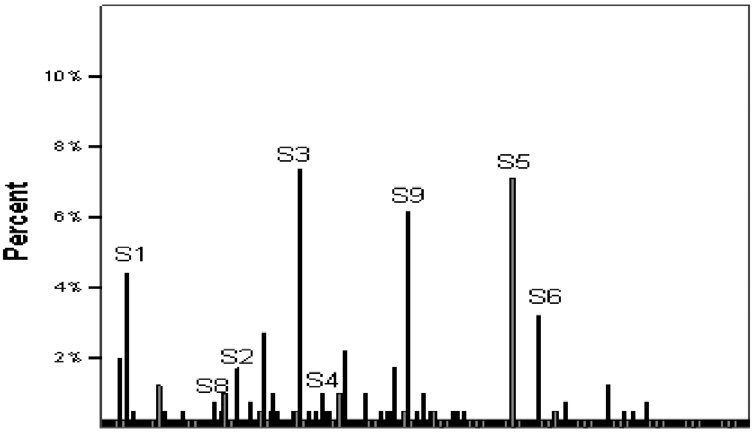


Figure 5.19: Distribution of suppliers in cluster 3.

The distribution of creators in cluster 3 shows a top eight. Only five of those eight came to light in the univariate models, taken all of them together. Not one model would have given the same results. Further, the univariate models brought creators forward that in cluster 3 are not that important, like C6, C3 and C5. Concerning the suppliers, a top four is presented in cluster 3, consisting of S1, S3, S9 and S5. Other outliers from the univariate models (S8, S2, S4 and S6) are pushed back. Again, not one model alone could have presented the same top five suppliers.

To have a view on where the top eight of creators of cluster 3 is situated in the three univariate models, we refer to Figure 5.20, 5.21 and 5.21. The top eight is marked by the letters A through H. In the small cluster of Model A, we find a top three (as already discussed), with two of the creators similar to the top eight of cluster 3 and one new outlier. The rest of the top eight is situated a bit lower in frequencies. But even in that range of frequencies, new creators are put forward in this model. The same situation is found for the outliers of Model B. Two creators of the top three are the same as in the top eight of cluster 3, except that it does not concern the same two creators. Here A and D are in the top three, while in Model A this was C and D. Again, in the lower frequency range the remaining six creators of the top eight were found, along with some new creators. In the small cluster of Model C, a top five creators presents itself, with four out of five coming from cluster 3's top eight. One has even an extreme frequency of 25.6% (again creator D). One of the top eight (C) has fallen very low in frequency, while new creators rise. Regarding the outliers in terms of creators, no univariate model shows the same results as the multivariate model.

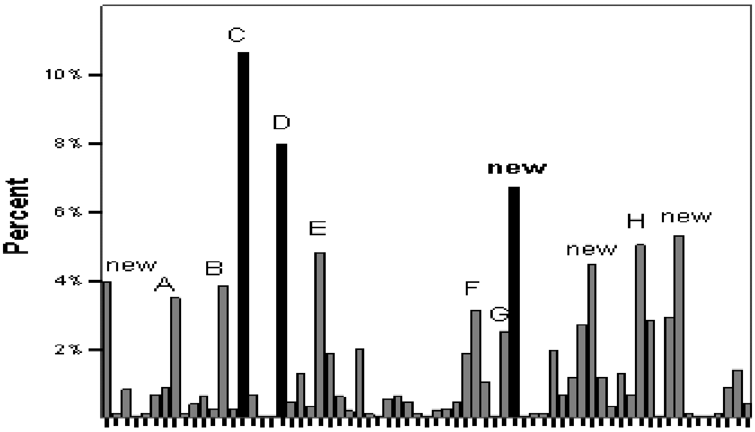


Figure 5.20: Location of top 8 creators cluster 3 (A-H) in Model A.

After situating the top eight of creators of cluster 3 in the univariate models, we do the same with the top four of suppliers. In Figure 5.23, 5.24 and 5.25 we mark this top four with letters W through Z. Looking at the distribution of suppliers in the small cluster of Model A, we had distinguished a top six, here marked in bold. Three of the top six are also to be found in the top four of cluster 3, and three new

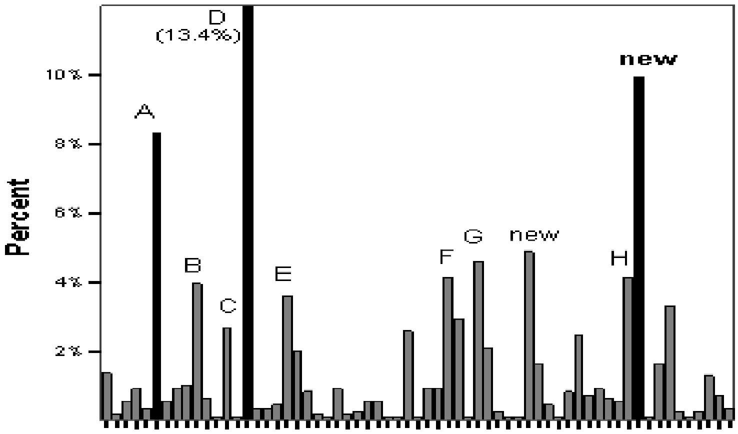


Figure 5.21: Location of top 8 creators cluster 3 (A-H) in Model B.

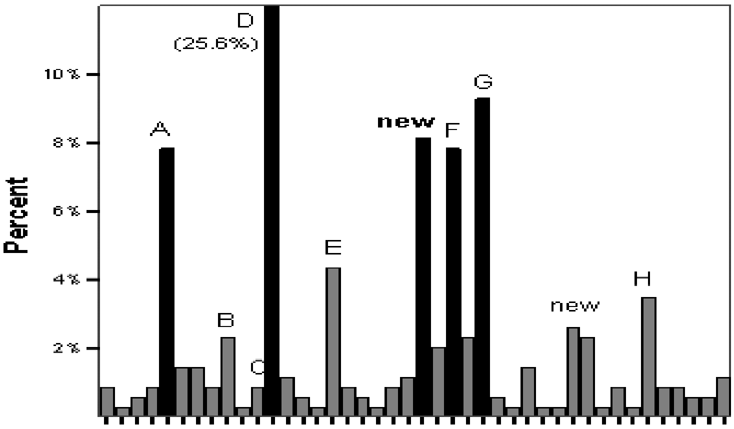


Figure 5.22: Location of top 8 creators cluster 3 (A-H) in Model C.

outliers are found. One of the top four of cluster 3 (Y) has fallen low. In Model B, we have a top three, with only one supplier similar to the top four of cluster 3, namely supplier Z. Not only are new suppliers put forward in the top three, but also in the frequency range where suppliers X and Y are to be found. Supplier W, a part of the top four in cluster 3, is in this model not even a peak on the graph. It has a frequency of .6% in this small cluster. The opposite is true for one of the two new outliers in this top three, which is not represented at all in cluster 3. At last, if we look at the suppliers that draw attention in the small cluster of Model C, we find a top three, with two of them also being part of the top four of cluster 3. However, at least five other suppliers outnumber supplier W and Z if it comes to frequency. Like the situation with the creators, also the suppliers could not be represented by a univariate model in the same way as by the multivariate model.

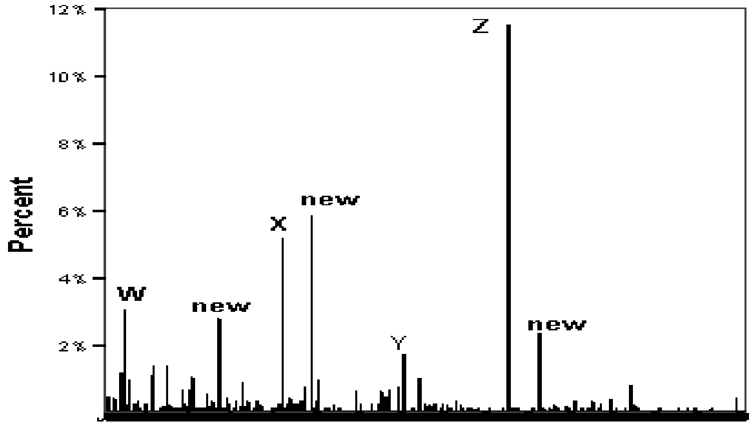


Figure 5.23: Location of top 4 suppliers cluster 3 (W-Z) in Model A.

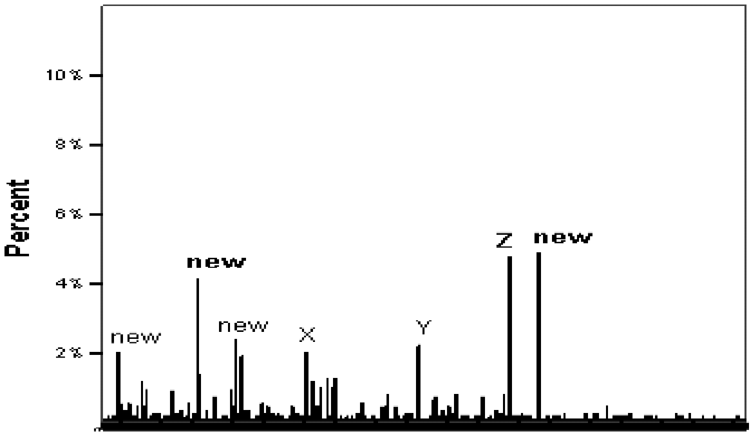


Figure 5.24: Location of top 4 suppliers cluster 3 (W-Z) in Model B.

All graphs comparing multivariate versus univariate clustering up till now concerned categorical attributes. One might argue however that the only cases we ac-

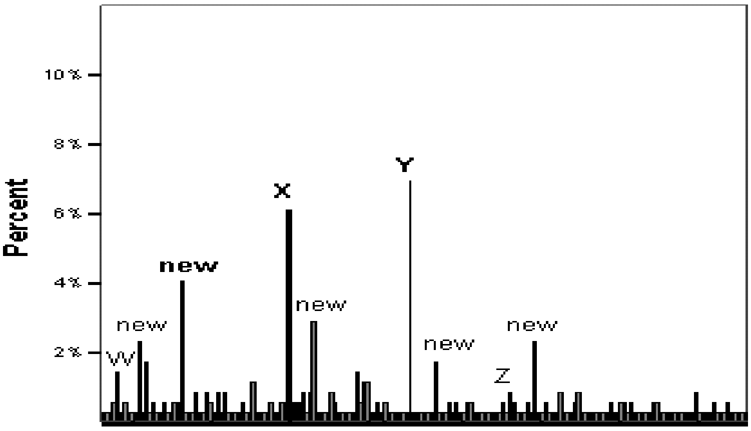


Figure 5.25: Location of top 4 suppliers cluster 3 (W-Z) in Model C.

tually visualize in the small cluster of the multivariate analysis, are PO's with high values for the numerical attributes. When looking at Figures 5.26 till 5.34, we see that this is not the case. For example, if we look at 5.26, we see that just taking all PO's with a higher value of number of changes (like model A does) does not deliver the same sample as what is selected in the small cluster of the multivariate model. For example: only a proportion of PO's with 13 changes is selected in the multivariate model, in contrast to Model A where all PO's with 13 changes are selected. This is reflected by the higher percentage of PO's with this number of changes in Model A. The same analysis can be made for Figure 5.28 where we see that even PO's without changes after release are incorporated in the small cluster of the multivariate model, in contrast to the univariate small cluster. Notice that Model C is trained on this attribute and that there is a clear split: PO's with 3 changes after release or more are categorized in the small cluster. All other Figures are to be interpreted the same way. Only the last numerical attribute, count over limit (Figure 5.35, 5.36, and 5.37), does not show a different composition of the various small clusters. This could suggest that the count over limit is not as important as the other numerical attributes in detecting cases with a high risk profile.

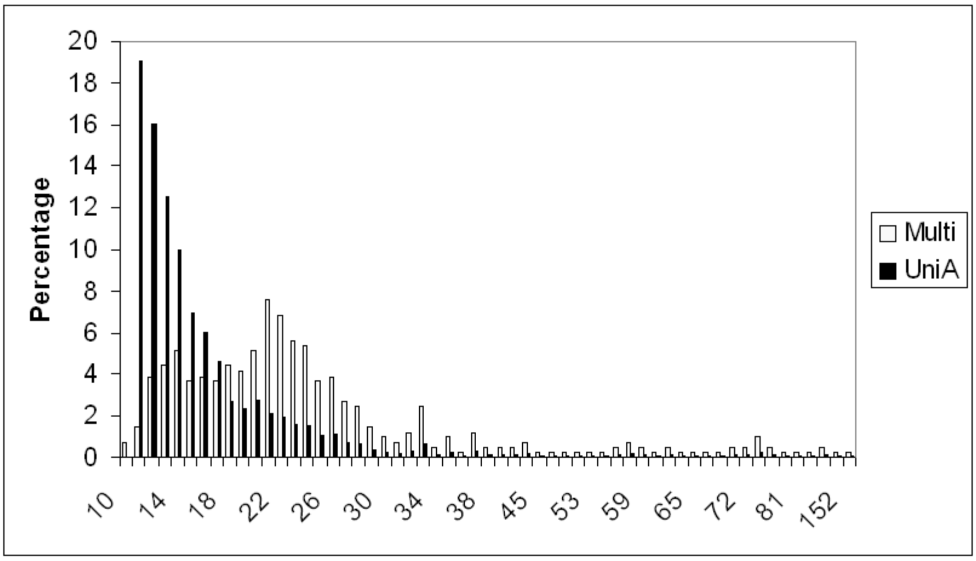


Figure 5.26: Distribution of 'Number of changes' in small cluster of Model A and of the multivariate model.

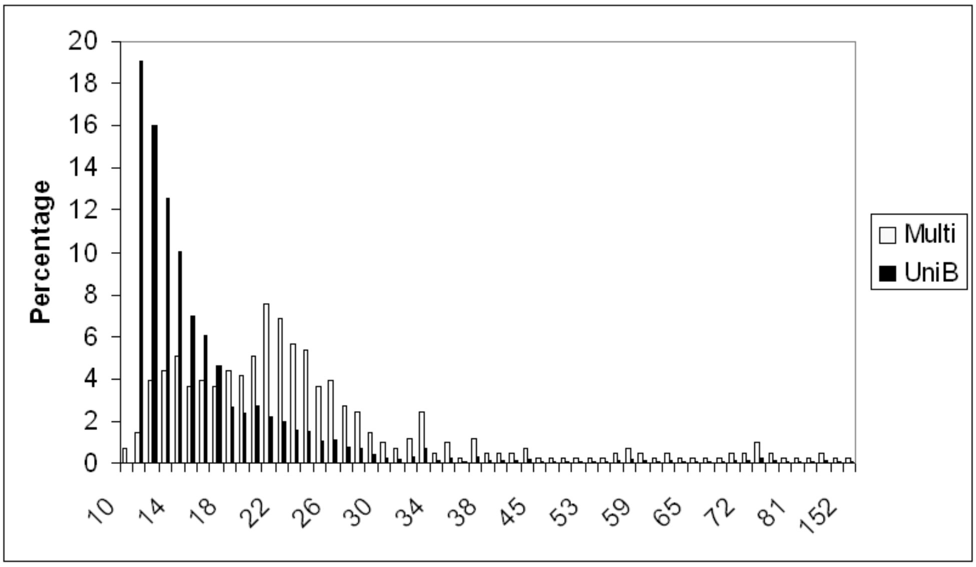


Figure 5.27: Distribution of 'Number of changes' in small cluster of Model B and of the multivariate model.

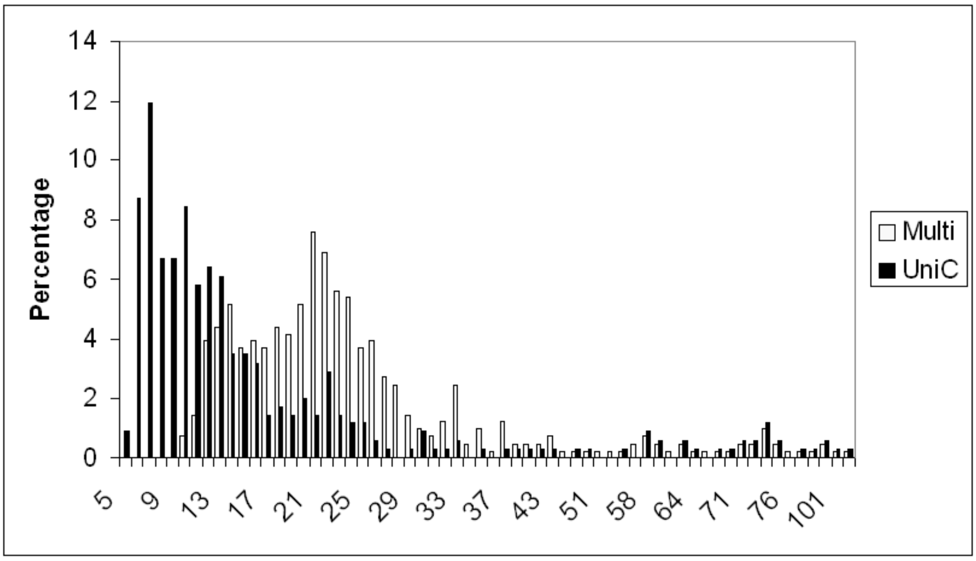


Figure 5.28: Distribution of 'Number of changes' in small cluster of Model C and of the multivariate model.

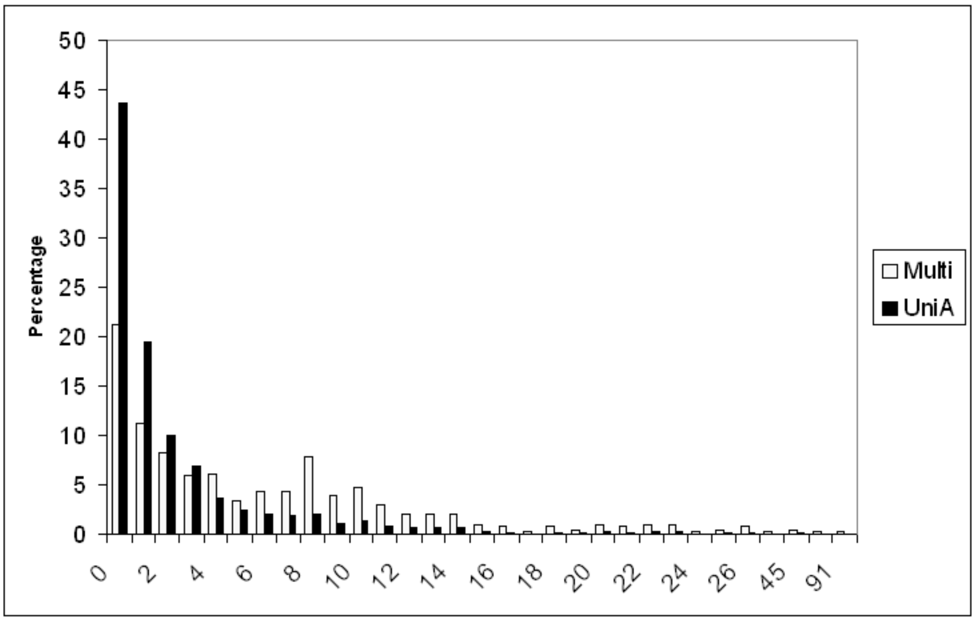


Figure 5.29: Distribution of 'Number of changes after release' in small cluster of Model A and of the multivariate model.

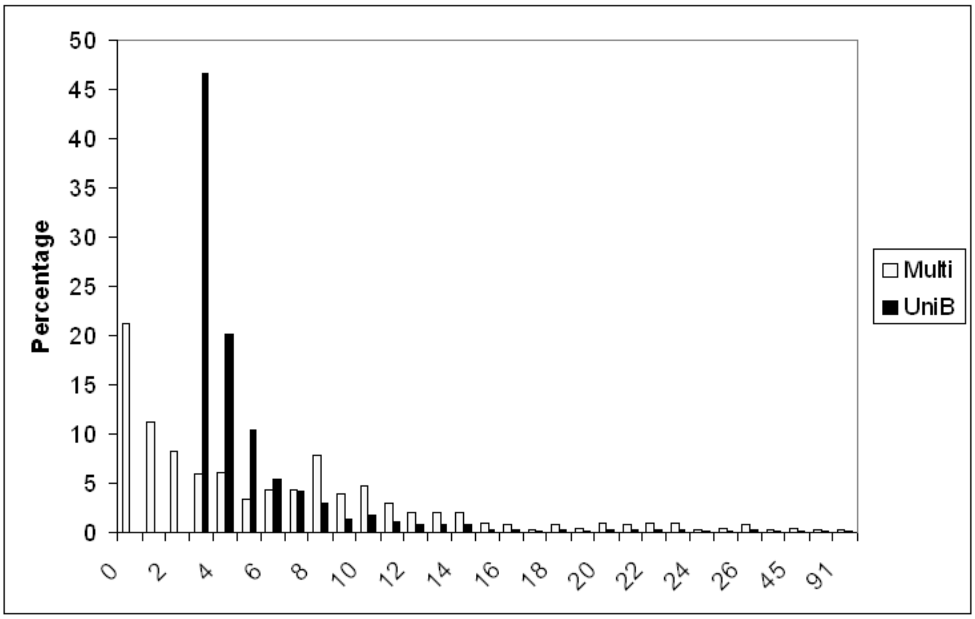


Figure 5.30: Distribution of 'Number of changes after release' in small cluster of Model B and of the multivariate model.

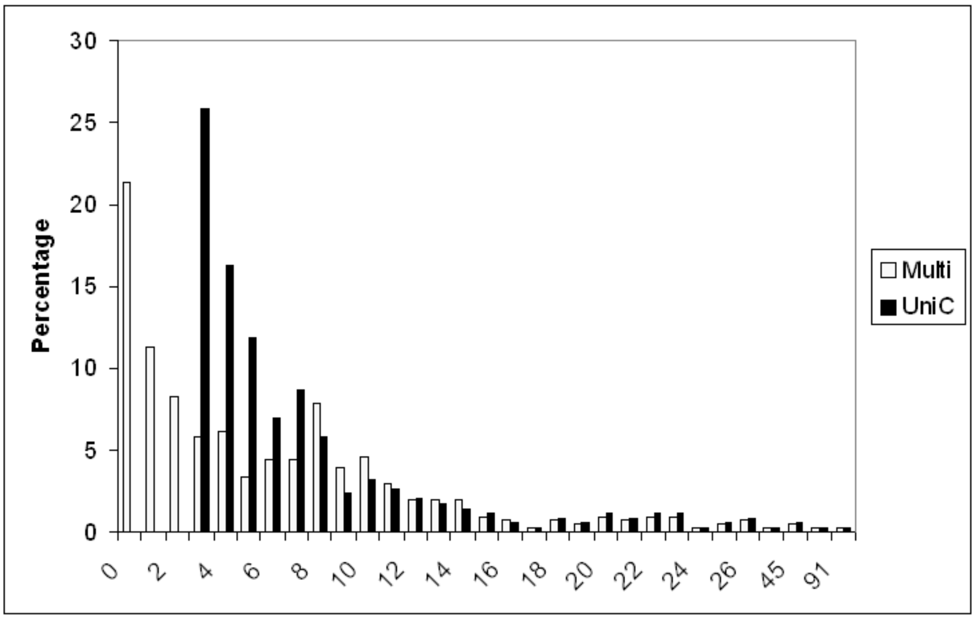


Figure 5.31: Distribution of 'Number of changes after release' in small cluster of Model C and of the multivariate model.

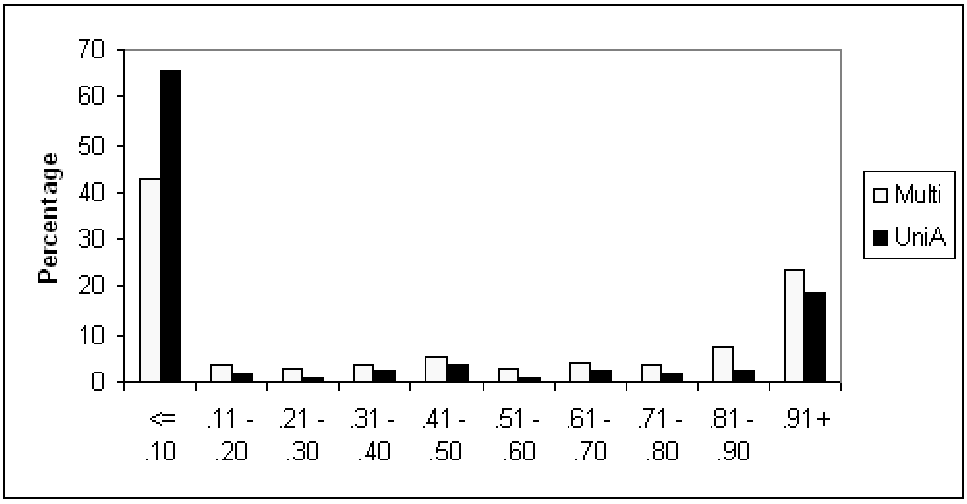


Figure 5.32: Distribution of 'Percentage price related changes after last release' in small cluster of Model A and of the multivariate model.

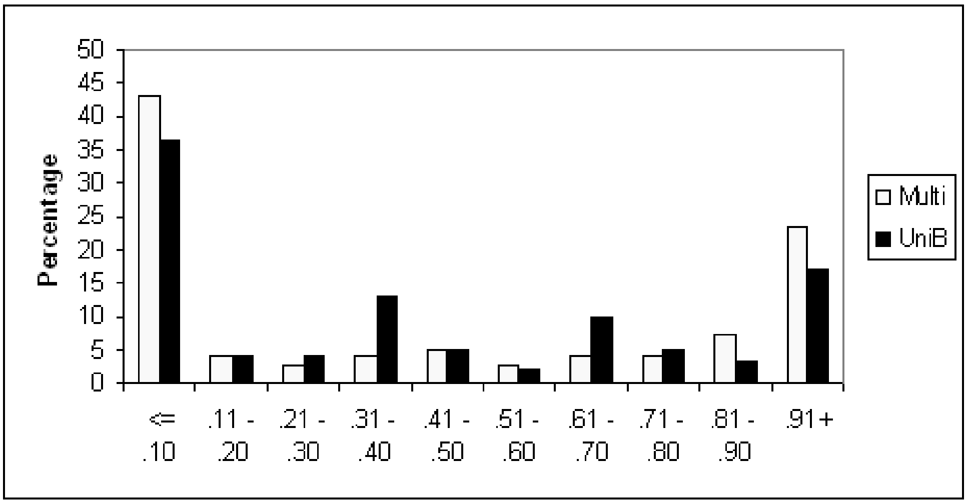


Figure 5.33: Distribution of 'Percentage price related changes after last release' in small cluster of Model B and of the multivariate model.

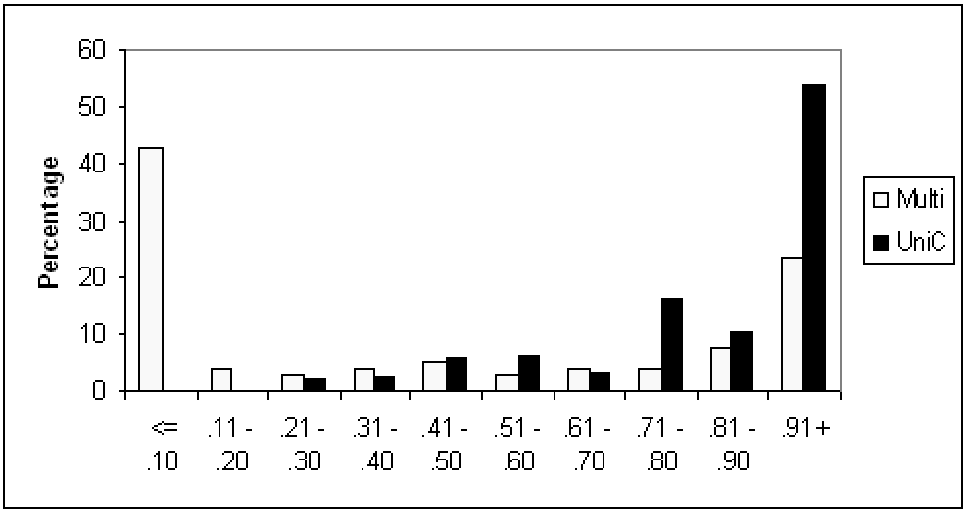


Figure 5.34: Distribution of 'Percentage price related changes after last release' in small cluster of Model C and of the multivariate model.

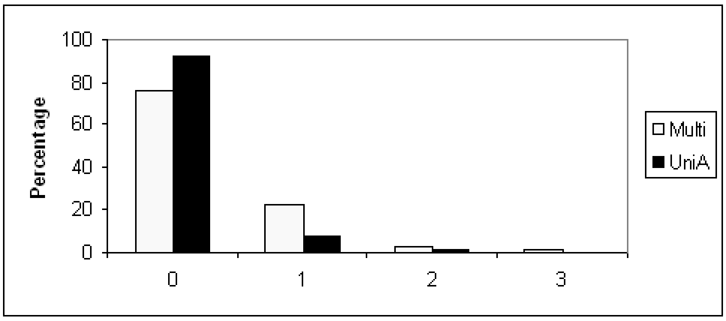


Figure 5.35: Distribution of 'Count over limit' in small cluster of Model A and of the multivariate model.

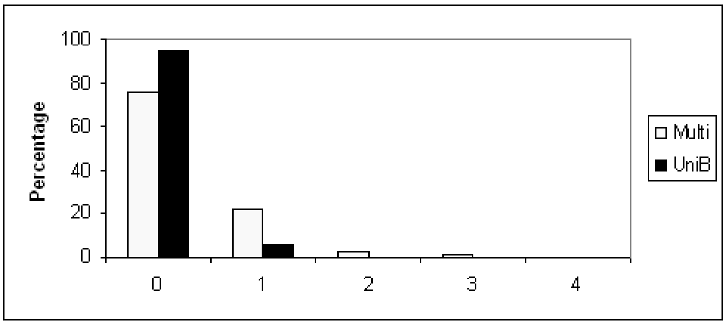


Figure 5.36: Distribution of 'Count over limit' in small cluster of Model B and of the multivariate model.

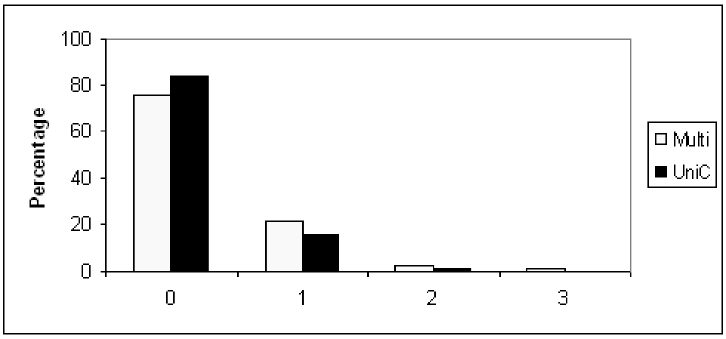


Figure 5.37: Distribution of 'Count over limit' in small cluster of Model C and of the multivariate model.

5.7 Conclusion

In this chapter we apply the IFR^2 Framework, presented in Chapter 2, in a top 20 ranked European financial institution. This framework suggests to implement a descriptive data mining approach for reducing internal fraud risk. A major difference with the current reporting techniques is that we use background knowledge to create behavior describing attributes, but that the algorithm turns this into knowledge. In this particular case, it was known that a lot of fraud risks were associated with changes conducted on PO's. This is used as input for the creation of attributes. After this step, the data mining approach took us further in our analyzes, as opposed to regular reporting techniques, where the analyst himself searches the data attributes for anomalies or interesting patterns.

The results of the case study suggest that the use of a descriptive data mining approach and the multivariate latent class clustering technique, can be of additional value to reduce the risk of internal fraud in a company. Also, using univariate latent class clustering did not yield the same results. We can further state that the application of the suggested methodology at the case company produced a tone of more concern about asset misappropriation in the procurement cycle.

Chapter 6

The Extended IFR^2 Framework

In Chapter 2, the IFR^2 Framework is introduced. The core of this framework is to apply a data mining approach in order to reduce internal fraud risk. In this chapter, an extension is added to the IFR^2 Framework, resulting in the 'Extended IFR^2 Framework'. In the extended IFR^2 Framework, the original IFR^2 Framework is complemented with a process mining part. Process mining is a relatively new research domain that aims at uncovering a process model based on real transaction logs. This can be applied in several ways for the purpose of internal fraud risk reduction.

We start this chapter with an introduction in process mining. In the following section the contributions of process mining to internal fraud risk reduction are given. Next, the Extended IFR^2 Framework is presented, followed by an explanation of how process mining can be seen as a complement to the IFR^2 Framework.

6.1 Process Mining

Nowadays many different information systems, like ERP, WFM, CRM and B2B systems, are characterized by the omnipresence of event logs. These can be referred to as 'audit trails', 'transaction logs', 'history' etcetera. Traditionally, an organization stores a lot of this kind of information as background information, but does not actively use this information to analyze the underlying process. This is where process

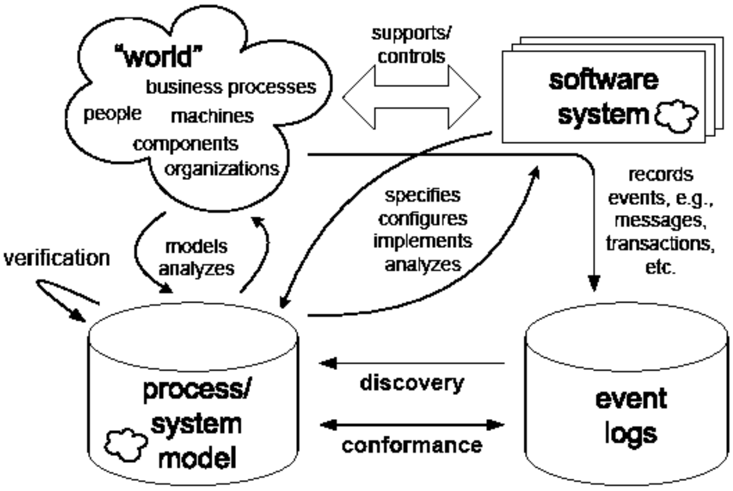


Figure 6.1: Concept of Process Mining

mining aims to make a difference. "*The basic idea of process mining is to diagnose processes by mining event logs for knowledge*" (van der Aalst and de Medeiros, 2005). In Figure 6.1 an illustration of the concept of process mining is given.

With process mining, several assumptions are made. First of all, one assumes it is possible to record events such that at least four characteristics can be identified. An event 1) refers to an *activity*, 2) refers to a *case*, or process instance, 3) can be appointed to a performer, or *originator*, and 4) a *timestamp* can be identified. For each process under investigation these are the constraining assumptions. If available data fulfills these assumptions, process mining can be applied on that particular process. Table 6.1 shows a classic example of an event log, used by van der Aalst et al. (2007), van Dongen et al. (2005) and van der Aalst and de Medeiros (2005) amongst others. The event log shows an example with 19 events, allocated to five cases, describing five different activities, performed by six persons.

Event logs are the starting point of process mining. The data of the event log can be mined and different aspects about the underlying process can be analyzed.

Table 6.1: An example of an event log, used by van der Aalst et al. (2007).

Case id	Activity id	Originator	Timestamp
case 1	activity A	John	9-3-2004:15.01
case 2	activity A	John	9-3-2004:15.12
case 3	activity A	Sue	9-3-2004:16.03
case 3	activity B	Carol	9-3-2004:16.07
case 1	activity B	Mike	9-3-2004:18.25
case 1	activity C	John	10-3-2004:9.23
case 2	activity C	Mike	10-3-2004:10.34
case 4	activity A	Sue	10-3-2004:10.35
case 2	activity B	John	10-3-2004:12.34
case 2	activity D	Pete	10-3-2004:12.50
case 5	activity A	Sue	10-3-2004:13.05
case 4	activity C	Carol	11-3-2004:10.12
case 1	activity D	Pete	11-3-2004:10.14
case 3	activity C	Sue	11-3-2004:10.44
case 3	activity D	Pete	11-3-2004:11.03
case 4	activity B	Sue	14-3-2004:11.18
case 5	activity E	Clare	17-3-2004:12.22
case 5	activity D	Clare	18-3-2004:14.34
case 4	activity D	Pete	19-3-2004:15.56
case 1, 3:	A-B-C-D		
case 2, 4:	A-C-B-D		
case 5:	A- E -D		

In general, three different perspectives can be distinguished: the process perspective, the organizational perspective and the case perspective. The *process perspective* or the "How?" question focuses on the ordering of activities. Which paths are followed? This will typically be expressed in terms of Petri Nets or Event-driven Process Chains (EPC). The *organizational perspective* or the "Who?" question uses the input data in the field 'Originator'. In this perspective, underlying relations between performers or between performers and tasks can be exposed. The *case perspective* or the "What?" question focuses on a single case. This will be more interesting to analyze if other data elements than in the event log are added in a separate table, for example the size of an order or the related supplier etcetera. (van der Aalst et al., 2007)

In the context of internal fraud risk reduction, all three perspectives are important, as will be pointed out later. In the next section we first start with the added value process mining can deliver to our framework. Afterwards we present the Extended IFR^2 Framework and discuss the three mining perspectives in the light of internal fraud risk reduction.

6.2 Contributions of Process Mining to the IFR^2 Framework

Process mining has several contributions to fraud risk reduction, being the reason why it is added to the IFR^2 Framework. A first contribution is its open mind set and objectivity. Other advantages are specifically related to the advantage process mining has over data mining. These advantages are expounded next, along with the actual difference between process and data mining. Further, the possible impact of process mining on monitoring internal controls and continuous control monitoring are highlighted.

6.2.1 Open Mind and Objectivity

One of the advantages of applying process mining is that, just like with the data mining application, it is not necessary to have a specific fraud in mind. When one has a specific fraud in mind when interpreting the analysis and looking if there are opportunities to commit this specific fraud, one can be blind for other opportunities.

Surplus value is delivered by the objectivity with which the process mining techniques work, without making any presuppositions. This feature is especially of value for the process perspective. When mining the organizational and the case perspective, it can indeed be beneficial to have some specific fraud(s) in mind. The advantage of an open minded start is even greater with process mining than with data mining, because the description of behavior with data mining (to create the input data) requires some background knowledge and pre-processing which also might influence the mining. Other advantages process mining delivers over data mining are described next.

6.2.2 Process Mining versus Data Mining

Process mining actually is data mining on an event log. Process mining aims at discovering patterns data, which makes it a subset of data mining. The main difference is the data to mine. At process mining, the input data is per definition an event log. The event log is a data set that describes the process followed, by means of ordering the process activities and providing attributes of them. Process mining aims at uncovering the underlying process (process perspective), laying bare underlying relationships (organizational perspective), and testing certain assertions (case perspective). We can conclude that process mining is a subsection of data mining, and hence more narrow than data mining. The assumption in the IFR^2 Framework when referring to descriptive data mining, is that the input data for the data mining tries to describe the behavior or the *process output data*. In the case of procurement as business process, the 'behavior of a PO' was summarized in the data that was mined. At the process mining leg of the Extended IFR^2 Framework, the *process data* itself is mined and hence patterns in the process are looked for.

By mining the process itself instead of the outcome of a process, several contributions are created. A first contribution is that we are working more at the root of the investigated business process. Accordingly, when a researcher is able to discover malfunctions within the process, this is far more efficient than finding all the deviating outcomes, resulting from this malfunctioning, and not knowing the underlying problem. A second contribution is that process mining can respond better to frauds resulting from collaboration between employees. The organizational perspective of process mining can uncover unusual collaborations between certain employees or more frequent collaborations than assumed to be. Caution is still required, as certain phony collaborations that are according procedures and that do not raise any questions, will still stay unnoticed. However, the organizational perspective is an extra contribution as opposed to the data mining leg. This issue also addresses the fourth element of the fraud diamond, the capability of a perpetrator to beat the system. A third contribution is the activity of monitoring internal control. This is the fifth component of COSO, for which process mining provides a real suggestion. The data mining aimed merely at the fourth component, control activities.

6.2.3 New Approach of Monitoring Internal Control

It was already stated that the IFR^2 Framework is an extension of the internal control framework. The IFR^2 Framework suggests an approach for the fourth component of the COSO framework, to implement control activities. In her framework, COSO identifies the monitoring of controls as the fifth component of internal control. Employees need to know that non-compliance with controls is likely to be detected (deterrence effect). Further, monitoring controls also provides feedback concerning these controls and. (Cosserat, 2004) Business practice found ways to deal with this fifth component. Mostly the monitoring of internal control is executed by means of reprocessing. This includes running transactions, designed by the auditor, on the information system, to test the system's ability to identify and correct errors in the capturing and processing of data. These tests are run to test password security, user identification codes and input editing controls. (Konrath, 2002) This approach of monitoring internal controls is further extended with some random sample testing. The process mining approach however suggests a completely new approach of monitoring internal controls.

The case perspective of process mining provides the possibility to monitor well specified internal controls, like for example the segregation of duty. Current business practice tests this segregation by a) testing a random sample on this assertion, and b) running a violation against this assertion. The process mining case perspective provides us with the possibility to test every case that passed the system on this assertion. In other words, we can test post ante whether the internal control settings worked effectively for all passed cases, instead of testing whether the internal control settings work correctly at that single moment in time (when the tests are run). The situation can be compared as follows: instead of checking whether a fence is open or closed at a specific moment (current method of internal control), we now have a look at what succeeded to pass the fence. This provides a whole new view on monitoring internal controls. Also, with process mining we are able to test all passed cases, as opposed to only a random sample.

In the context of monitoring internal controls, it is also appropriate to link this to the concept of continuous monitoring, a specific area of control monitoring. Although we wish to introduce process mining in a broader context than continuous monitoring alone, in the following paragraph the concepts of continuous auditing and continuous monitoring are shortly introduced.

6.2.4 Continuous Auditing and Monitoring

A new age of data storage goes along with new demands. Traditionally, internal audits and their related testing of controls are executed on a cyclical basis. However, with the electronic storage of all kinds of data, easily accessible and available in large volumes, new methods of internal auditing are implemented. Advanced technology has been employed to perform continuous auditing. Continuous auditing is defined as *"a framework for issuing audit reports simultaneously with, or a short period of time after, the occurrence of the relevant events"*. (CICA/AICPA, 1999) An important subset of continuous auditing is the continuous monitoring of business process controls (Alles et al., 2006). Continuous monitoring of controls is defined by the Institute of Internal Auditors as *"a process that management puts in place to ensure that its policies and procedures are adhered to, and that business processes are operating effectively. Continuous monitoring typically involves automated continuous testing of all transactions within a given business process area against a suite of control rules."* (IIA, 2005) Notice that continuous monitoring is a responsibility management bears, while continuous auditing is a task of the internal audit department. However, there is an interaction effect between the efforts put into place concerning continuous monitoring and continuous auditing. When management performs continuous monitoring on a comprehensive basis, the internal audit department can partly rely on this and no longer needs to perform the same detailed techniques as it otherwise would have under continuous auditing. (IIA, 2005)

Aside from the contributions process mining has in the context of internal control monitoring, there is also a contribution to further investigate in continuous monitoring. Through process mining, continuous monitoring can be applied on a frequent basis (monthly or weekly).

6.3 The Extended IFR^2 Framework

The way we suggest to extend the IFR^2 Framework with a process mining part, is visualized by Figure 6.2. The first steps remain the same. It is only after the transformation of data, with input of domain knowledge, that process mining comes into the framework. Although process mining and data mining are visualized on equal height and hence can be executed in both follow order or parallel to each other, the process mining part can also be seen as an exploratory investigation before turning to the data mining part of the framework. This is visualized by the arrow of 'Process Mining' toward 'Transformation of data'.

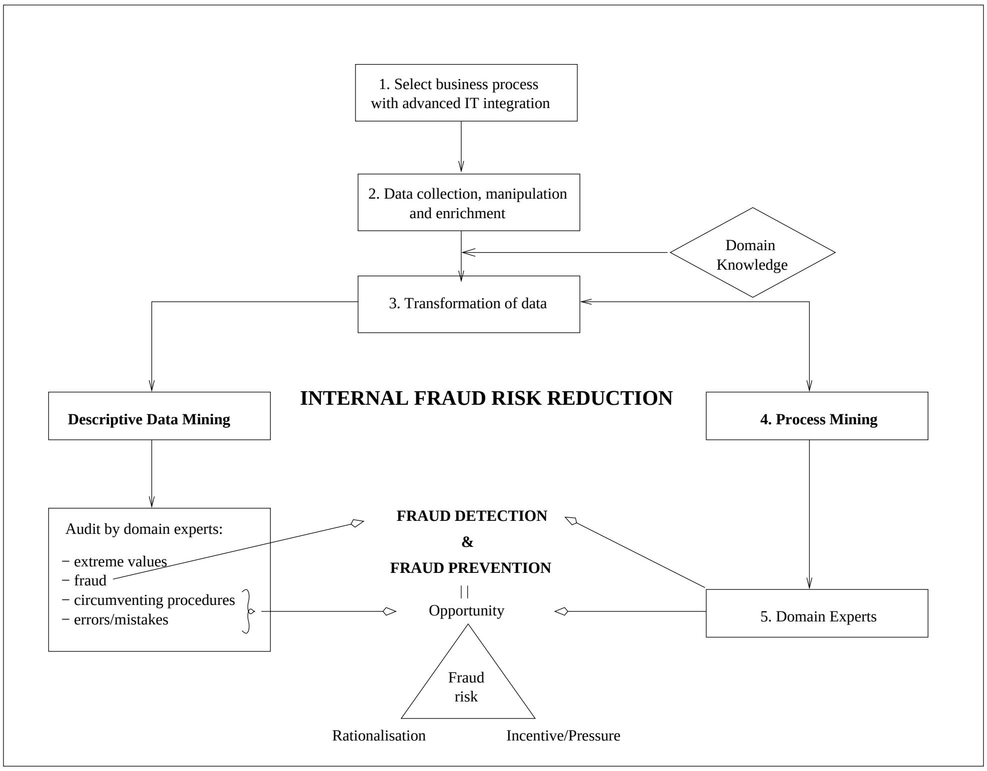


Figure 6.2: The Extended IFR^2 Framework

After applying process mining, feedback from and to the domain experts is needed to interpret the results. This will eventually lead to new insights whether or not there are opportunities to commit fraud. The interest in opportunities in terms of fraud prevention stays the same as in the data mining part, as explained in Section 2.3. Finding cases where certain assertions not hold on the other hand, is part of fraud detection. Also, information gathered from this process mining step can be implemented in a data mining approach and lead indirectly to fraud detection.

6.4 Process Mining for Internal Fraud Risk Reduction

In this section the three previously mentioned process mining perspectives are reviewed in terms of internal fraud risk reduction. Which perspective serves which end and how does it contribute to the IFR^2 Framework?

6.4.1 Process Perspective

The process perspective of process mining is the most natural perspective to begin with. This perspective uses the *activity* information of an event, and the order of the activities as they occur per process instance.

Usually, an organization has business processes mapped out in procedures, guidelines, user guides etcetera. In a first process mining step, we perform a conformance check: is the event log conform to the (designed) process model? The event log of a real business process will usually contain lots of different patterns (a specific combination and order of activities on one process instance). In order to perform a conformance check, it is advised to separate the most frequent patterns of the unfrequent patterns. The *pareto*-ratio can be a valuable aid in this situation. This ratio states that around 20% of the patterns will be able to describe 80% of the event log. Taking the patterns of this run of the mill is a good way of reducing the number of found patterns to a realistic number of patterns to discuss with the domain experts.

Supplementary to this conformance check, we compare the actual process with the designed process, enabling a *Delta analysis*, i.e. detecting discrepancies between the process design constructed in the design phase and the actual execution in the enactment phase (van der Aalst et al., 2003). This kind of analysis is important in the light of defining opportunities to commit fraud. In this manner one can detect flows or sub flows that for example were not meant to exist. This can in turn give insights in potential ways of misusing or abusing the system.

6.4.2 Organizational Perspective

At the organizational perspective, using mainly the *originator* information, it is possible to get a view on the activities originators perform. It is for instance possible to look at a task-by-originator matrix or to create a visual map with all connections between originators. These outputs are the input for a discussion with the domain expert. Again, a risk assessment by opportunity check is the main goal of this discussion.

The value of the organizational perspective in terms of internal fraud risk reduction depends on the type and structure of the company and business process. If there is no -or less- strict structure about who works with whom or who performs which tasks, this perspective can obviously not add as much value as if these structures are set more strict.

6.4.3 Case Perspective

The case perspective of process mining focuses on a lot more information than the activities and their order (process perspective) or the originators (organizational perspective). The case perspective takes into account very different attributes. These are mostly stored in additional tables. These attributes can concern the process instance itself, or a particular activity. For instance: when the process under investigation is the procurement process and the process instances are invoices, attributes about the process instance (the invoice) can be the invoice number or the invoice value, while attributes of an activity can contain information about that particular activity. The typical activity attributes which are available in the event log are Originator and Timestamp (who performed this activity and when?), but a lot more information can be stored. For instance, of the activity 'post the invoice at the general ledger' information like on what account this invoice is posted and with which reference number, is information that can be kept as extra attributes. These extra attributes are stored in an additional table.

As the name suggests, mining the case perspective of a log, looks at the event log case by case. This makes it possible to check certain aspects or conditions which should be met on a case basis. If one takes the procurement business process for instance, one person may have the authority to create a Purchase Order and another person to approve the following invoice. This is a control on the transactional level (two different persons have the authorities for two different tasks). It can however occur that one person has both authorities (both to create a Purchase Order and to approve an invoice). On itself, this has not to pose any problem, as long as that person does not enhance these two authorities on one single case. Therefor it would be interesting to enforce a segregation of duties on the case level instead of on the transactional level. The case perspective of process mining can check, case by case, whether this segregation of duties is respected or not. As a matter of facts different explicit controls, such as the segregation of duties for example, can be checked at a case base level.

A very valuable feature of this case perspective is that it allows us (and in a later instance the company) to test a complete data set on specific conditions, for example on a monthly base. This is a huge improvement compared to the random checks currently performed in the context of monitoring internal controls.

In Chapter 8, we turn to the implementation of this Extended IFR^2 Framework in Epsilon. But first we elaborate on the creation of the event log, used as input for our case study.

Chapter 7

Event Log Creation

In Chapter 6 we introduce the concept of process mining and incorporate this in our framework. For the application of process mining, one needs an event log to mine the process. The data for such an event log can be provided by all kinds of information systems. Also at Epsilon, one can find trails of actions amongst others in the ERP system, SAP. We are in particular interested in trails or logs relating to the procurement process, because this is the business process under investigation. In Chapter 4, an overview of the available data is already presented. In this Chapter an overview is provided of how exactly this data is converted into a usable event log for our study.

7.1 Events in the Procurement Process

An important assumption at process mining is that it is possible to describe the process under consideration by sequentially recording events. These events are the activities that all together constitute the process. After discussing the procurement process with the domain expert, we arrive to the process model depicted in Figure 7.1.

For modeling the described process we use a Petri Net. A Petri Net is a dynamic structure that consists of a set of *transitions*, *places* and *directed arcs* that connect these transitions and places in a bipartite manner. Transitions are indicated by boxes and relate to some task, while places are indicated by circles and represent passive phases. Places may hold one or more *tokens*, indicated by black dots. If all input places of a transition contain a token, this transition is *enabled* and may *fire*. When a transition fires, it consumes a token of each of the input places and produces a token

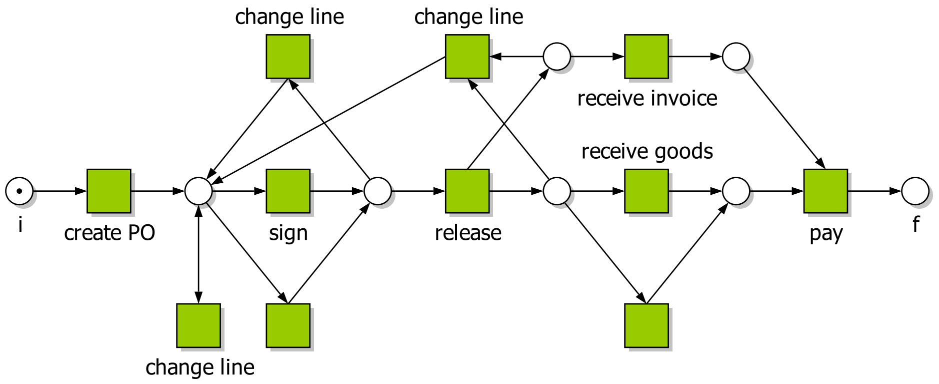


Figure 7.1: Process model of procurement in Petri Net

for each of its output places. (see e.g. Reisig (1985))

The Petri Net in Figure 7.1 represents the procurement process at the case company. After the creation of the PO and the item line, the PO is released. Depending on the PO, an additional signature (sign) could be needed before it can be released. Often, the item line will be changed between the creation and the Sign and Release activities. It is also possible that the item line is changed after it was released, and hence a new sign and release need to be triggered. Only after a release, eventually goods and an invoice are received, in any order. After receiving the invoice and goods, a payment can be made. Normally, both a Goods Receipt and Invoice Receipt are prerequisites. However, in some circumstances no Goods Receipt is necessary. In these cases the Goods Receipt Indicator must be turned off.

Aside from the possibility to determine the process by sequential events, it is also assumed that these events are all linked to one particular case, called a *process instance*. We must ask ourselves *'What would be a correct process instance to allocate events to?'* The answer on this question is extremely important in process mining, because this determines the event log substantially.

A natural choice of process instance would be a PO, since this seems to be the central document where everything relates to. But do we have data available to link all steps to a PO and to construct as such audit trails per process instance, being a PO? The answer is short: yes, we have this information. We know exactly who made a PO; who signed and released it and when; we know when the Goods and Invoice Receipts are obtained and by whom; and we know when these invoices are paid. Still, we cannot use a PO as process instance. This rejection is on grounds of the dynamics a PO can experience. We know for example which PO is signed or released, we do not know however anything about the content of the PO at that time. This means that we can see for example that a PO has been signed and released for ten times, but we do not know the exact content of what has been approved each time. The same holds for the related Goods and Invoice Receipts. We know there is a link, but we do not know if the content of the invoice was for example also part of the PO when it was signed and released. These lacunae are created by the double dimensionality used in saving and linking data. An invoice line is for example matched with a line item of a PO. This is also the base of the ERP system to control the approval. So a line item could be a better candidate for process instance.

In this paragraph we examine the feasibility of using a PO item line as process instance. Is there a technical link between a PO item line and a sign, a release, a Goods Receipt, an Invoice Receipt, and payment? For most of them the answer is straightforward 'yes', but sometimes it is only indirectly possible to place a link. Looking at the data available and the process depicted in Figure 7.1, we consider the PO item line as the best candidate for process instance. We established the following events as activities of the process:

- Creation of the PO (parent of item line)
- (Change of the particular item line)
- Sign of parent PO
- Release of parent PO
- Goods Receipt on item line (GR)
- Invoice Receipt on item line (IR)
- Payment (or Reversal) of item line

These events are also called Work Flow Model Elements (WFMElt). In what follows, we work out how the process instances were collected and selected and how each WFMElt is created. In the end, the event log contains per process instance different events, being a *WFMElt*, with a particular *Timestamp* and *Originator* for each event.

Table 7.1: Model example of event log

PI-ID	WFMElt	Event Type	Timestamp	Originator
450000000190	Create PO	Complete	02 Feb 2006	John
450000000190	Change Line	Complete	30 Nov 2006	John
450000000190	Sign	Complete	05 Dec 2006	Paul
450000000190	Release	Complete	06 Dec 2006	Anne
450000000190	GR	Complete	05 Jan 2007	John
450000000190	IR	Complete	15 Jan 2007	Matt
450000000190	Pay	Complete	16 Feb 2007	Marianne
450000000210	Create PO	Complete	23 Jan 2007	Doug
...				

Also the *Event Type* must be stated, but this will be set default to ’Complete’, since we do not have information to distinguish further. In Table 7.1 a model event log is given. Of course, the event log based on real life data will not look as clean.

Partly motivated and inspired by the data structure within SAP, but also partly arbitrarily, an item line of a PO is selected as process instance. This decision is however susceptible to comments. An alternative event log could be made by starting from invoice item lines. That way also invoices without related PO make part of the investigation. Limited in time and resources though, this dissertation only takes a PO item line as process instance.

7.1.1 Process Instance ID

Taking a PO item line as process instance, we constructed the process instance ID (PI-ID) by compressing the PO number with the line item number. At Epsilon, the PO number is a 10-figures number, starting with 45, generally put as ’45xxxxxxxx’. The line items are mostly numbered 10, 20, etc. The resulting process instance IDs will have a format of 45xxxxxxxx$\underbrace{\textit{itemline}}$. For example the first PI-ID in Table 7.1 refers to PO 4500000001, item line 90.

7.1.2 Selection of Process Instances

The process instances that are subject of the event log creation are the PO item lines that are incorporated in the table EKPO during 2007. All of these are namely item lines where invoices of 2007 are attached to. This gives us a good starting point for creating an event log of process instances that covered the whole process.

7.1.3 Creation of parent PO

Although we selected an item line of a PO as process instance, the start of the process still lies in the creation of the PO that parents this item line. For this event, we retrieve the timestamp and originator from table EKKO, where the header information of each PO is captured. The 'date of creation' and 'employee ID' give us the desired information. The link between the process instance and EKKO has been made by the PO number.

7.1.4 Change Line

The PI-IDs included in the previous table 'Creation of parent PO' are the basis of the event log. For more information about changes we have to switch to CDHDR and CDPOS. Only at the item level of changes (contained in CDPOS) we can make a link between the change and the PO item line. So first, CDPOS was enriched with header information from CDHDR, such as 'data' and 'employee ID'. Next, an inner join between this table and the 'Creation of parent PO' table on equal PI-IDs was conducted. To make sure we only select item line changes and no schedule or assignment line changes, the extra condition of 'table name = EKPO' is added.

7.1.5 Sign and Release

For this WFMElt, we actually skip to a different level than the process instance's level. The process instance is on PO item line level, while a 'Sign' or 'Release' occurs at the level of a PO as a whole. When a PO is signed or released this information is only captured on one location, namely in the change logs (CDHDR and CDPOS). To retrieve the correct timestamp and originator we conduct a left join of 'Create PO' and CDHDR on equal PO numbers and select all approvals (transaction code = ME28). To distinguish between a sign and a release, we turn to CDPOS and look at the field that is changed (field name) and what value it turned in to (value new). Whenever the field FRGKE has a new value ’5’, ’9’, ’A’ or ’I’ it concerns a sign, whenever the new value is ’6’, ’8’, ’C’ or ’J’ it concerns a release.

7.1.6 Goods Receipt and Invoice Receipt

For the next two WFMElts we are back to the process instance’s level since a Goods and an Invoice Receipt occur at the level of an item line. The information about when this receipt is introduced into the ERP system and by whom, is captured in table EKBE. We can search the correct timestamp and originator on PO number and PO line item number. The ’category’ field in EKBE tells us whether we are dealing with a Goods Receipt (GR: category = E) or an Invoice Receipt (IR: category = Q). How the selection of GRs and IRs is carried out, will be explained in the next paragraph, after we explained which data is used for the ’Pay’ WFMElt.

7.1.7 Payment or Reversal of PO item line

At the tables BKPF (header information) and BSEG (item information) we find all kinds of information about financial documents. The ’FD numbers’ starting from 500.000 (till 1.000.000) are reserved for invoices. We joined the header and item information for these numbers, extracted the related PI-ID (to be found in BSEG, after compressing PO number and item line number), selected ’employee ID’ (BKPF) and ’clearing date’ (BSEG) as originator and timestamp respectively. If the field ’reverse doc nr’ is empty, this means the item line was paid and WFMElt is set ’Pay’. If the field ’reverse doc nr’ however contains a document number, this means the invoice was not paid, but reversed by a credit note. In this case, WFMElt is put ’Reverse’. The program is written as just described, but no ’Reverse’ is present in the event log.

7.1.8 Selection of Goods Receipt and Invoice Receipt

To create the event log as clean as possible, we took precautions about the selection of IRs and GRs. To make sure we only have relevant receipts, we carried out a backward selection. We started from the payments we introduced to our event log and worked our way back to the related goods receipt, via the invoice receipt. The connection between the payment of an item line and the receipt of the invoice in question, has been made through the 'object key' in BKPF, which refers to the document number in EKBE (compressed with the material document year). So after creating a set of PI-IDs with all possible IRs, we took an inner join between the PI-IDs with a payment and the PI-IDs at the IR-table, resulting in only the IRs followed by a payment.

We undertook the same procedure for the selection of GRs. This time the connection had to be made between the (selected) IRs and the GRs. The 'ref. doc.' provided the desired information. Again, an inner join was taken of the table of IRs and GRs on attribute 'ref. doc.' being the same.

7.1.9 Critical Remarks

As already mentioned, the decision to take the item line of a PO as process instance is influencing the event log and hence the process mining analysis. Other selections of process instances could be investigated later in future research.

Another criticism can be given on the signs and releases. These activities occur at PO header level and as a result are not imperative related to the item line under investigations. Perhaps a change on another item line triggered a new Sign/Release. This also is to be brought back to the double dimensionality used in SAP, which we cannot completely discard.

A last criticism can be given on the originator of the WFMElt 'Pay'. The stated 'Originator' is in fact the person that introduces the invoice into the ERP system. The fact that it gets paid is because of the approvals on the PO. So strictly spoken, the stated originator at 'Pay' or 'Reverse' in the event log is the person entering the invoice that will be paid or reversed afterwards. We are however mainly interested in the activity 'Pay' or 'Reverse' itself, so this limitation does not pose a problem for our research.

7.2 Data Attributes

Aside from the information at the event log, being WFMElt, Originator and Timestamp, other attributes of a PI are interesting to record. We created an additional table with data attributes of each PI. This table contains information about the par-

Table 7.2: Example of data attributes of PIs

PI-ID	Name	Value
450000000190	Doc Type	DI
450000000190	PG	B01
450000000190	Supplier	12345
450000000190	Net Value	10.000
450000000190	Unit	EA
450000000190	Quantity PO	1
450000000190	GR Ind	X
450000000190	GR Total Quantity	1
450000000190	GR Total Value	10.000
450000000190	IR Total Quantity	1
450000000190	IR Total Value	10.000
450000000190	Pay Total Value	10.000
450000000210	Doc Type	FO
...		

ent PO and about the item line itself. More specifically, the following information was recorded about the parent PO: the document type, the purchasing group that creates this PO and the supplier involved.

The information we selected about the process instances concerns the value (in euros) of the item line (Net Value), the unit in which the quantity is expressed (Unit), the amount of ordered units on this line (Quantity PO) and whether or not the GR indicator was flagged (GR Ind). If this indicator is flagged, the input of a GR is mandatory for the payment of the invoice. If it is turned off, an invoice can be paid without a GR. Next to this PO related information, we also included the total quantity and total value of all Goods Receipts that are linked to this PO item line (GR Total Quantity and GR Total Value). We did the same for the related Invoice Receipts (IR Total Quantity and IR Total Value) and the total value of all Payments that are associated with this process instance (Pay Total Value). In Table 7.2 an example of the recorded data attributes of a process instance is shown.

Besides data attributes of PI's, also a table with extra attributes of audit trail entries is created. In particular four events are enriched with additional information:

‘Change Line’, ‘IR’, ‘GR’, and ‘Pay’. If the event concerns a ‘Change Line’, we store information about the change: if it was a change of the net value, what was the absolute size of this modification? If not the net value was changed, but another field, for example the delivery address, this field contains a modification of zero. The other stored attribute gives us, in case of a change in net value, the relative size of the modification (hence a percentage). If the event concerns an ‘IR’, four attributes are stored. We store the references that contain the (possible) link to the ‘GR’, the link to ‘Pay’, the quantity of the units invoiced, and the credited amount, the value. Notice that these quantities and values only concern this specific Invoice Receipt, as opposed to the Invoice Receipt related attributes of the Process Instance. The latter attributes provide summarized information of all Invoice Receipts attached to the Process Instance. Also beware that this information is not collected from an entire invoice, but only from the specific line that refers to the PO item line of this process instance. Similar to the ‘IR’, three attributes are stored when the event concerns a ‘GR’: the reference to possibly link this Goods Receipt to the associated ‘IR’ (this is not always possible, only in a specific number of cases), the quantity of goods received and the resulting value that is assigned to this Goods Receipt. This value is the result of multiplying the Goods Receipt quantity with the price per unit agreed upon in the PO. The last event that is provided of attributes is ‘Pay’. The value of this payment is captured, as well as the key to create a link to an associated ‘IR’. Table 7.3 summarizes all attributes created on both the process instance and the audit trail entries level.

7.3 Cleaning the Event Log

One last step is made in the context of cleaning the event log. The way the log is constructed now, not all audit trails end with ’Pay’, like it is preferred to be. (one ending activity facilitates the process mining, and ’Pay’ is the last activity in the process design) This could be solved by adding an artificial ’End’ task before we start mining this process. However, we choose to clean up the event log in such a way that all cases start with ’Create PO’ and end with ’Pay’. There are two ways we can obtain this. Either we filter out all process instances that do not end with ’Pay’, or we keep all selected process instances, but cut off the audit trail after the last ’Pay’ activity of that trail. The latter option is chosen. This choice is inspired by the fact that if we filter out all PO’s that do not end with ’Pay’, we might filter out a certain group

Table 7.3: Attributes of event log

Level	Attribute	WFMElt
Process Instance	Document type	
	Purchasing Group	
	Supplier	
	Order Quantity	
	Order Unit	
	Net Value	
	Goods Receipt Indicator	
	IR Total Quantity	
	IR Total Value	
	GR Total Quantity	
	GR Total Value	
	Pay Total Value	
Audit Trail Entry	Modification	Change Line
	Relative Modification	Change Line
	Reference GR	IR
	Reference Pay	IR
	Quantity IR	IR
	Value IR	IR
	Reference IR	GR
	Quantity GR	GR
	Value GR	GR
	Reference IR	Pay
	Value	Pay

of PO's that behave in a different manner. Think for example of PO's that are being used over and over again. The audit trail of such a PO may look as follows: *Create PO-Sign-Release-GR-IR-Pay-Change Line-Sign-Release-GR-IR-Pay-Change Line-...* By filtering PO's on 'end task equals 'Pay' ' we could create a bias on the proportion of this kind of PO's in the total data set. By cutting off the audit trail after the last payment, we preserve the original representation of PO behavior.

We now turn to Chapter 8 for the analysis of this event log.

Chapter 8

Process Mining Application

In this chapter, we implement the Extended IFR^2 Framework at Epsilon by mining the event log of Chapter 7. The process mined, is the process of entering PO's into the ERP system and its following activities until payment.

In the first part of the process mining case study, we want to support the ideas of the domain experts about the process and if needed to reveal weaknesses in the process. For this purpose, we perform a process diagnostic step in Section 8.1. A good methodology for process diagnostics by process mining can be found in Bozkaya et al. (2009), which is the applied methodology in this study. It consists of five phases: log preparation, log inspection, control flow analysis, performance analysis, and role analysis.

Process diagnostics only focuses on a global view of the business process, in order to help the analysts and domain experts to reveal weaknesses and problems in the business process. This step can be compared with the descriptive statistics in econometrical analyzes or with univariate data mining results. The goal is to have global view on the process. In the second part of this case study, we advance to a verification step. In this step, we check whether certain aspects and assertions of the process are held or not. This verification uses the case perspective of process mining and will be elaborated in Section 8.2. Because this verification step would have extreme added value when applicable on a complete data set instead of on a random sample, we perform this second step also on the procurement data of January 2007 in Section 8.4, after a short process diagnostics step on January in Section 8.3. But first we start our analysis on a random sample out of the whole year 2007 to have a general view. We start with the five phases of process diagnostics on this random sample.

8.1 Process Diagnostics

The methodology of Bozkaya et al. (2009) for process diagnostics by process mining, encompasses five phases: log preparation, log inspection, control flow analysis, performance analysis, and role analysis. We start with the log preparation.

8.1.1 Log Preparation

In the previous chapter we gave a detailed overview of the log preparation. One last step, after the cleaning step, is conducted before analyzing the event log. For reasons of computability, a random sample of 10,000 process instances out of 402,108 was taken. After collecting all the data necessary for the event log, ProM*Import* is used to convert our event log into the desired MXML format. The MXML file is analyzed using the open-source tool ProM. (van der Aalst et al., 2007) ProM consists of many different algorithms to cover the analysis of the three process mining perspectives and is designed in such a way that researchers and users can easily develop their own plugins and add them to the framework. For more information about ProM, we refer to van Dongen et al. (2005) and to `www.processmining.org`.

8.1.2 Log Inspection

We start this second phase with a log summary, providing some descriptives of the log. Next, a short introduction to the *FuzzyMiner* plugin, used for further inspection, is given. At last, the results of the log inspection by means of the *FuzzyMiner* are given.

Log summary

As already stated, we start with a random sample event log of 10,000 process instances. A process instance is a PO item line. The process analyzed in this study contains seven real activities (see Table 8.1). Notice that the event 'Reverse' does

Table 8.1: Log events

WFMElt	Occurrences (absolute)		Occurrences (relative)	
	original log	cleaned log	original	cleaned log
Pay	11,157	11,157	17.842%	18.123%
IR	10,648	10,608	17.028%	17.231%
Release	10,651	10,471	17.033%	17.009%
Create PO	10,000	10,000	15.992%	16.244%
Sign	9,794	9,616	15.663%	15.62%
GR	5,235	5,213	8.372%	8.468%
Change Line	5,045	4,497	8.068%	7.305%

not occur in this log.[1] The cleaned log at hand contains 61,562 events in total and 285 originators participated in the process execution[2]. All audit trails start with the event 'Create PO', but originally they did not all end with 'Pay'. The ending log events were 'Pay' (93.85%), 'Change Line' (5.02%), 'Release', 'IR', 'GR' and 'Sign'. As described in the previous chapter, a cleaning step is conducted such that all audit trails end with 'Pay'. The occurrences of the audit trail entries of both the original and the cleaned log can be found in Table 8.1. As can be seen all 'Pay' activities are maintained, and there are still 10,000 process instances involved (because every audit trail starts with 'Create PO'). The log summary confirms that all audit trails end with the activity 'Pay'. This cleaned log will be our process mining input.

To get a first glance on the real process, we start mining the process perspective of the log. This typically results in a graphical representation of the process underneath the transactions of the event log. To construct such a process model, causal dependencies have to be exposed. The causal dependency between activity A and activity B, denoted by $A \rightarrow_W B$, means that A is directly followed by B, but B is not directly followed by A. This dependency is deducted by looking at the (timed) order of activities per process instance (or case). Looking at the classic process mining example in Table 6.1, showed in Chapter 6, we see that $A \rightarrow_W B$, $A \rightarrow_W C$, $A \rightarrow_W E$, $B \rightarrow_W D$, $C \rightarrow_W D$, and $E \rightarrow_W D$. In this model example these dependencies are easy to deduce, but in real life logs, this is not only harder, there are also two

[1] 'Reverse' is apparently not present at all in the log for Belgium (not even before random sampling).

[2] The original log contained 62,530 events and 290 originators were involved.

important complicating factors: noise and completeness.

For a causal dependency between A and B, B must follow A directly, but A may never follow B. So if an event log contains for example 99 out of 100 times the sequence A - B, and one time B - A, it will not give the dependency $A \rightarrow_W B$. So we need a mining algorithm that can handle *noise*.

Further, *completeness* can be a problem. If there are n activities that can be executed in parallel, the total number of possible causal dependencies is $n!$, growing faster than an exponential function. Hence it is not realistic that every log contains all paths possible for the underlying process. This leads us to a supplementary condition: the mining algorithm needs to be able to handle low frequent behavior.

The *FuzzyMiner* plugin is a good algorithm to tackle the problems of noise and completeness. Also the *HeuristicsMiner* is a good alternative. Because the *FuzzyMiner* is better able to clean up a large amount of confusing behavior (typical to real life logs), the *FuzzyMiner* is preferred above the *HeuristicsMiner*. In what follows a short description of the *FuzzyMiner* approach is given.

Fuzzy Mining

The resulting graph of a flow-mining step (process perspective) represents all activity classes as a node and all precedence relations as arcs. With simple processes, this does not pose any problem, but with less-structured or complex processes, this visualization can result in a very chaotic graph. The developers of the *FuzzyMiner* identified the challenges of a process mining technique as the challenges of the field of cartography: to simplify highly complex and unstructured topologies. ”Activities in a process can be related to locations in a topology and precedence relations to traffic connection.” (Gunther and van der Aalst, 2007) Using the concept of a roadmap as a methaphore: imagine a map that depicts every city and town in the same way and makes no distinction between an interstate or a simple road. The map could be perfectly correct, but not suitable. The same holds for the result of process mining. Process mining techniques need to be able to provide a high-level view on the process, hereby abstracting from undesired details. (Gunther and van der Aalst, 2007)

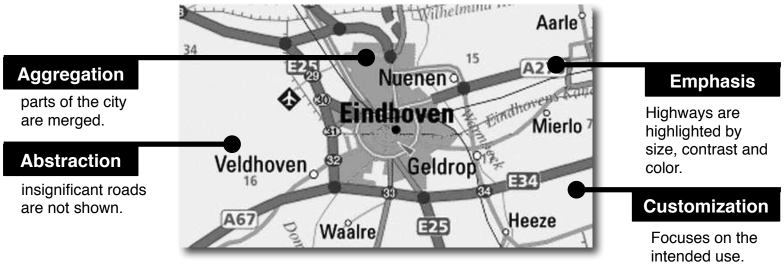

Aggregation
parts of the city are merged.
Abstraction
insignificant roads are not shown.
Emphasis
Highways are highlighted by size, contrast and color.
Customization
Focuses on the intended use.
Aarle
Nuenen
Eindhoven
Mierlo
Veldhoven
Geldrop
Waalre
Heeze
E25
A2
E34
A67

Figure 8.1: Example of a road map. (Source: Gunther and van der Aalst (2007))

Gunther and van der Aalst (2007) identified the solutions cartography came up with, and subsequently explored how these could be used to simplify and properly visualize complex and less-structured processes. The solutions of cartography can be bundled to Aggregation, Abstraction, Emphasis and Customization, depicted in Figure 8.1.

- **Aggregation:** Maps often visualize clusters of low-level information in an aggregated fashion, most often this is the case with (larger) cities. (e.g., the city Eindhoven in Figure 8.1)

- **Abstraction:** Information of less importance, in the chosen context, is abstracted from the visualization.

- **Emphasis:** Information of higher importance is emphasized visually, for instance by means of color, saturation and size.

- **Customization:** Depending on the local context, maps are specialized towards the use and purpose of the map.

Because these concepts are universal, well-understood, and established, Gunther and van der Aalst (2007) explored how they could be used to simplify and properly visualize complex, less-structured processes. In order to succeed in this, the authors had to develop appropriate decision criteria on which to base the simplification and visualization of process models. Two fundamental metrics were identified to support these decisions: (1) *significance* and (2) *correlation*.

Two types of significance metrics are used in the *FuzzyMiner* plugin: unary and binary significance metrics. The unary significance metrics evaluate the relative importance of a node (representing an activity) in the process model. The binary metrics evaluate the relative importance of an edge (representing a precedence relation). Within the correlation metrics, there is only one type used to simplify the model: binary correlation metrics, measuring how closely related two events, following each other, are. For the three types of metrics, different concrete metrics are selected and inserted in the *FuzzyMiner* plugin. More detailed information can be found in Gunther and van der Aalst (2007). We now turn to the inspection of the log.

Log inspection

For a first inspection of the event log, we run the *FuzzyMiner* algorithm with default settings. The result, depicted in Figure 8.2, reveals *Create PO-Sign-Release-IR-Pay* as the most frequent path. This is corresponding to the designed process model. Also the side paths are well explicable. The digress onto *Change Line* and the use of a Goods Receipt before the Invoice Receipt are part of the designed model. Also the path of having a Goods Receipt after a payment is easy to understand in light of a split delivery.

8.1.3 Control Flow Analysis

After this second phase of process diagnostics, log inspection, we turn to the control flow analysis, the third phase. In a first approach, we wish to uncover the core process that is embedded in the event log, and to be able to confirm that the business process functions in a way that corresponds to the designed model. In a second approach we wish to expose less frequent flows.

Uncovering the core process

Using the *Performance Sequence Analysis* plugin of ProM, we have a view on the patterns followed in this log. The analysis reveals 170 patterns. This is a very high

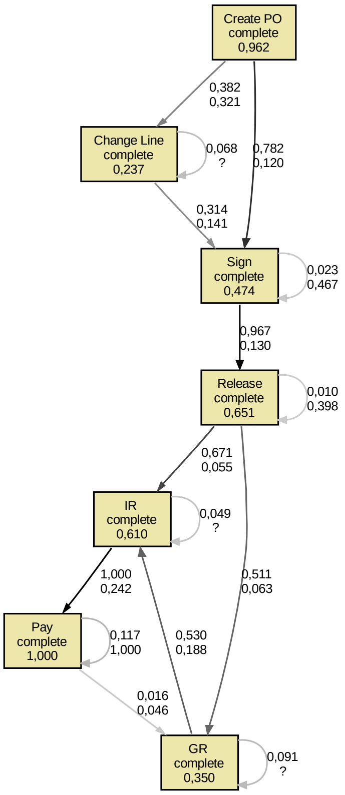


Figure 8.2: Result of *FuzzyMiner* with default settings

number, certainly for such a relatively simple process model design. This gives us already an idea of the complexity of this process and the 'noise' in this event log. Five, respectively eight patterns suffice to cover 78% and 91% of the entire log (see Table 8.2). Inspection of these patterns with the domain expert tells us already that all these patterns are completely according to Epsilon's procedures. Also the descriptive statistics about the occurrences and the throughput time reveal interesting information. This kind of analysis is part of process reporting, as described in Chapter 2.

To discover a process model that covers the run of the mill, it is necessary to filter out the infrequent patterns. That is why we will only use the first five patterns (describing 78% of the log) to discover a process model. This way we can extract an understandable process model from the event log that describes the overall process. This model will in turn be compared with the designed process model, in order to assure the process in general is executed as desired.

Apart from the most frequent patterns shown in Table 8.2, we also have a look at the most infrequent patterns, shown in Table 8.3. In this table we can read for instance that patterns 27 and 28 both occur only ten times (out of the 10,000). We see that 143 patterns are used to describe only 3.4% of the data set, while eight patterns sufficed to describe 91% of the data set. This fact confirms what we deduced before: there is a lot of noise on this data set. The reason why so many patterns are only followed a few times in the entire log, is because these patterns contain a lot of activities, with a maximum of 202 activities in one pattern. If even only one activity from a pattern misses, or switches place in such a trail, another pattern is formed. So the more activities a pattern counts, the less likely it is followed frequently exactly the same way. But though these numbers are interesting to see, these infrequent patterns still form a subset too large to examine manually. We now turn back to the core process in order to uncover the general process, followed at Epsilon. The infrequent patterns are filtered out in this step as outliers are discarded to run regression analyzes.

Table 8.2: Top 8 of most occurring sequences Random Sample

		Occurrences		Total	Throughput Time (days)			
Pattern	Sequence	#	%	%	Average	Min	Max	St.dev.
0	*Create PO - Sign - Release - IR - Pay*	3,066	30.7%	31%	16.75	3	176	9.59
1	*Create PO - Sign - Release - GR - IR - Pay*	2,048	20.5%	51%	35.95	2	327	26.33
2	*Create PO - Change Line - Sign - Release - GR - IR - Pay*	1.393	13.9%	65%	29.74	4	328	36.46
3	*Create PO - Change Line - Release - IR - Pay*	636	6.4%	71%	70.34	4	269	39.04
4	*Create PO - Change Line - Sign - Release - IR - Pay*	633	6.3%	78%	25.28	3	241	27.15
5	*Create PO - Sign - Release - IR - GR - Pay*	555	5.6%	83%	35.68	2	263	34.09
6	*Create PO - Sign - Release - Change Line - IR - Pay*	546	5.5%	89%	21.4	9	299	16.2
7	*Create PO - Release - IR - Pay*	232	2.3%	91%	20.04	2	197	24.16

Table 8.3: Most infrequent sequences

Pattern(s)	Occurrences per pattern	Total occurrences	Representation of log
27 - 28	10	20	0.2%
29 - 31	9	27	0.3%
32 - 35	8	32	0.3%
36 - 38	7	21	0.2%
39 - 41	6	18	0.2%
42 - 48	5	35	0.4%
49 - 58	4	40	0.4%
59 - 66	3	21	0.2%
67 - 91	2	50	0.5%
92 - 169	1	78	0.8%
		342	**3.4%**

Taking the selection of the log with only the five most occurring patterns and applying the *Final State Machine* (FSM) miner, results in a transition diagram which was the input for the tool *Petrify* to get a process model. The resulting process model is depicted in Figure 8.3. [3] Running a conformance check reveals that 80% of the total log (not only patterns 0-4), is covered by this process model. This result is used as a feedback to the domain experts. It was concluded that the general outlines of the process are clearly coming forward in the event log. This is seen as a reassuring start.

[3] The advantage of this model over the *FuzzyMiner* or *HeuristicsMiner* results, is the transition diagram that is straightforward to read. This is in contrast to the latter results, where one always has to consult the AND-OR semantics to know what exactly is depicted. This will be explained further in this chapter.

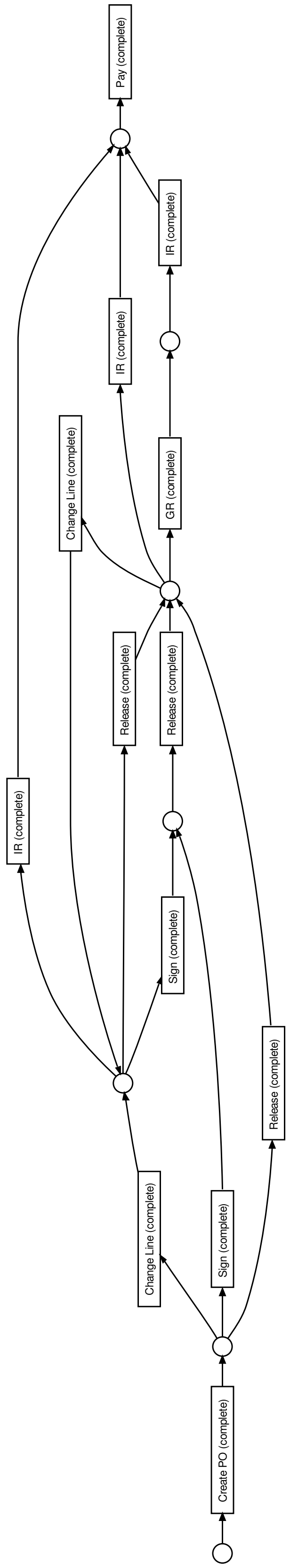


Figure 8.3: *FSM Miner* result of patterns 0-4

Exposing less frequent flows

Another contribution of the control flow analysis is to use the complete random sample with all 10,000 cases and to have a look at the resulting flows when lower thresholds are used. Lowering the threshold settings will result in a graph with more edges, exposing precedence relations that are less strong. This is a convenient way (visual) of looking at the most important infrequent paths. Turning back to the application of the *FuzzyMiner*, we change the settings in such a manner that more flows become apparent. Concretely, we change the 'Cutoff' edge filter to the values 0.70 and 0.85. These different settings indeed result in models with more edges. Elevating the 'Cutoff' to 0.70 (compared to the default setting of 0.20) revealed two extra flows. Elevating the 'Cutoff' further to 0.85 (depicted in Figure 8.4) revealed even four more extra flows (on top of the other two). The six extra flows are:

- Create PO → Release
- Sign → GR
- Create PO → GR
- Release → Pay
- Sign → IR
- GR → Change Line

Before discussing the extra six flows, visible at the graph in Figure 8.4, an important aspect of interpreting these results has to be highlighted. The arcs from one event to another in a resulting graph of the *FuzzyMiner*, need to be seen in an AND/OR relationship, which is not visible at this output graph. This means that for instance an arc from activity A to activity B does not per definition mean that B directly follows A. Perhaps this arc should be interpreted along with another arc, from activity A to activity C. The two flows 'A → B' and 'A → C' *may* represent an AND (or OR) relationship (after A, B and/or C follow) without having B per definition directly after A, the same for C. Hence, looking at the *FuzzyMiner* result gives us ideas of extra flows, but deducing direct flows between one activity and another, needs to be explicitly checked.

In the next paragraphs the six extra flows are discussed (based on interviews with the domain experts) and if necessary explicitly checked. Two flows are very normal: 'Create PO → Release' and 'GR → Change Line'. A 'Change Line' can occur at every stage of the process and the fact that the PO is not signed first, before it is released is a realistic possibility. However, there are certain conditions attached to

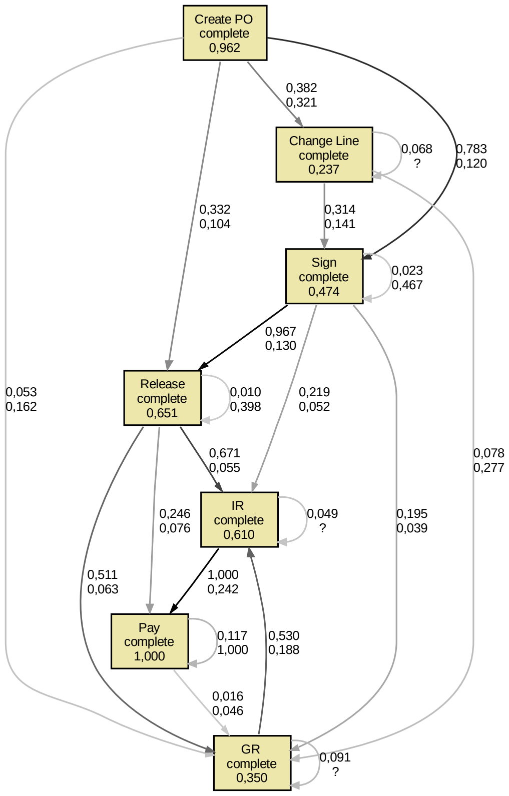


Figure 8.4: *FuzzyMiner* result with 'Cutoff'=0.85

leaving out the 'Sign'. To test whether there exists a direct flow between 'Create PO' and 'Release', we run the formula 'Eventually activity A next B' with A and B being 'Create PO' and 'Release' respectively at the *LTL-Checker* plugin. This reveals 258 direct flows from 'Create PO' to 'Release'. The presence of such flows highlights the necessity to later verify whether the procedures of leaving out the 'Sign' activity are followed. This will be checked at the verification step.

The flows 'Create PO → GR', 'Sign → GR' and 'Sign → IR' each have the same problem. A release is a prerequisite for ordering goods at a supplier (hence the name). Normally speaking, only after placing an order at a supplier, a Goods Receipt or an Invoice Receipt could be received at a purchasing department. Following this train of thought, all three flows are contrary to the designed process, and should also not exist if the SAP settings function as they should. Therefore, before looking for explanations or going over to investigation, we need to confirm whether these flows really occur in these specific orders, or that they are part of an AND/OR relationship that takes care of the above mentioned restriction. For this, we use the *LTL-Checker* plugin in ProM and test whether ‘Eventually activity A next B’ with A and B being the events in question we want to check. The LTL checks reveal that out of the 10.000 process instances, none showed the direct flow of ’Create PO → GR’, three instances had a direct flow ’Sign → GR’ and again none had a flow ‘Sign → IR’.

We take the three process instances with the flow ’Sign → GR’ under investigation. The first one shows a pattern of *Create PO - Sign - Release - Sign - GR - Release* . So a release took place before the Goods Receipt is entered into the system, confirming the SAP control settings. Because the events ’Sign’ and ’Release’ are both on the header level of a PO and not per definition linked to the process instance (item level), it could be that the ’GR’ in this case fell in between a *Sign - Release* flow, triggered by another line item. The other two process instances we looked into showed the same situation.

The last flow of the six, ’Release → Pay’, raises the question whether for these payments an according invoice is received. Normally, each ’Pay’ should be preceded by an ’IR’. Again we start with checking whether there exists a direct flow of ’Release → Pay’ for our process instances. We check the same formula ’Eventually activity A next B’ with A and B being ’Release’ and ’Pay’. There are 55 instances (i.e. only 0,55% of the cases) that show this direct flow. There are two possible scenarios for this flow: (1) the ’IR’ has taken place before ’Release’. This can again be explained as the ’Sign → GR’ flow: a *Sign - Release* flow, triggered by another line item, popped in between an *IR - Pay* flow of this process instance. Or (2), there is no ’IR’ related to this ’Pay’. If this is the case, there still could be a reasonable explanation, namely that another document than an IR is used, for example a ’Subsequent Debit’. These documents are not taken into account in our event log, because they are rarely used. The condition whether there exists an ’IR’ for each ’Pay’ will be tested and looked into later, at the verification step.

Table 8.4: Patterns of sequences with throughput time less than four days

		Occurrences	Throughput Time (days)			
Pattern	Sequence	(#)	Average	Min	Max	St.dev.
0	*Create PO - Sign - Release - IR - Pay*	153	3.01	3	3.96	0.08
1	*Create PO - Change Line - Sign - Release - IR - Pay*	80	3	3	3	0
2	*Create PO - Sign - Release - GR - IR - Pay*	4	2.25	2	3	0.5
3	*Create PO - Release - GR - IR - Pay*	4	2.75	2	3	0.5
4	*Create PO - Release - IR - Pay*	1	2	2	2	0

8.1.4 Performance Analysis

At the phase of performance analysis, questions like “Are there any bottlenecks in the process?” are answered. In this phase the average and maximum throughput times of cases are looked into and analyzed. It could be the case that a certain group of tasks is executed within a very short period of time, which might indicate an attempt to fraud. By means of example, all sequences that had a throughput time of less than four days are filtered out. This resulted in five patterns, described in Table 8.4. These five patterns represent 242 process instances, all with the end of their audit trail (the payment) being within four days after the creation of the parent PO. It is interesting to investigate these results, but as this kind of analysis is part of analysis based on (process) reporting techniques, this falls beyond the scope of the process mining application. However, this is a nice example of the complementarity between reporting techniques and data mining.

8.1.5 Role Analysis

At the fifth phase of process diagnostics, role analysis, the roles in a process are analyzed. A role should be seen as a person (in this case study) that is involved in the process by executing activities of that process. Role analysis attempts to answer questions like “Who executes what activities?” and ”Who is working with whom?”. In this phase, it is interesting to check on the efficiency of the segregation of duty.

The segregation of duty is a principle to reduce potential damage from the actions of one employee. (Elsas, 2008) Therefore it is hindered that one single employee has control over a critical combination of business transactions, such as there are for example a ‘Sign’ and a ‘Release’ authority in one single purchase. By looking at the role-activity matrix, we have a first look whether a person executing the activity ‘Sign’, also executes the activity ‘Release’. A print screen of a part of the matrix can be found in Figure 8.5. At this print screen we find for instance one originator that executed 1,733 times a release, and also signed 1,512 PO’s. This is the most extreme case of the event log, but other originators also combine these two tasks. This matrix does not tell us something about whether this should be a problem or not, because if these signs and releases concern different POs, there is nothing wrong with having both authorities in one person. We find however confirmation for the necessity to investigate this further. Also evidence is found about originators combining the activities ’GR’ and ’IR’ and ’Release’ and ’GR’. Here the same reasoning holds: in se this has not to be a problem, but again a further investigation is desired. These investigations require a case perspective of process mining, which brings us to the verification step, the second part of our analysis.

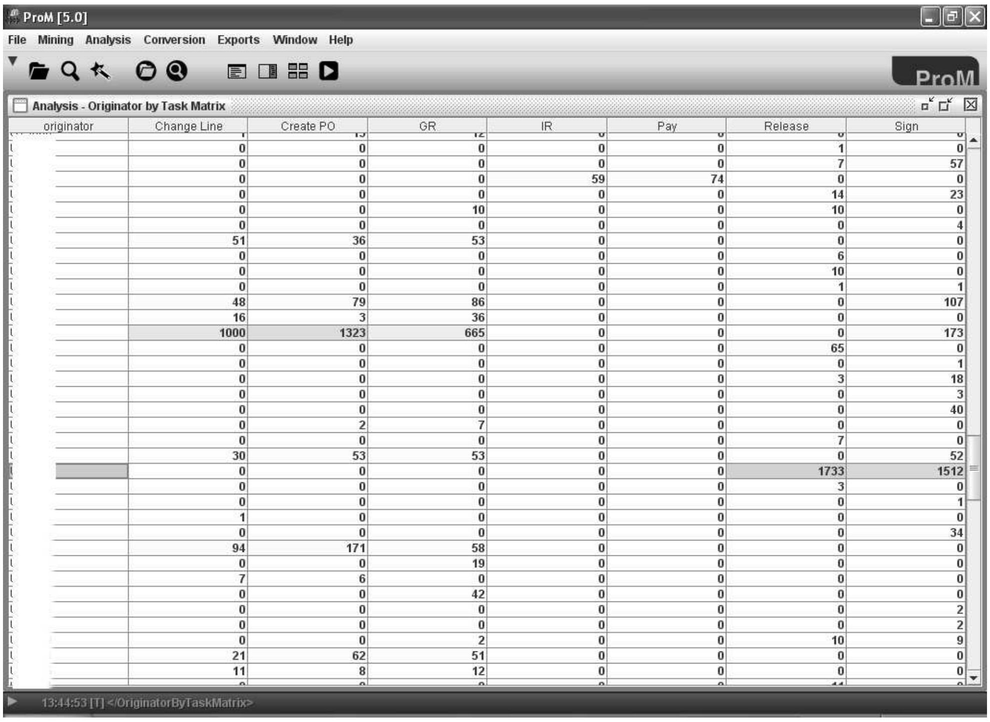
8.2 Verification Step

After mainly looking at the *process perspective*, we turn to the *case perspective* of process mining by the verification of certain properties. In this section we want to test whether certain assertions hold, i.e. whether the associated internal controls efficiently function. Several internal control settings are possible at an ERP environment. Rather than just checking whether these settings are in place at a specific moment, we test the output data on whether the internal controls function properly. We classify the checks to execute in three categories: checks on the segregation of duty, case specific checks and other internal control checks, not belonging to one of the two categories mentioned before. For all these checks, we use the *LTL Checker* plugin of ProM.

8.2.1 Checks on Segregation of Duty

As already confirmed by the role-activity matrix, there is a need for further investigating whether the segregation of duty is respected in this business process. After discussing with the domain expert which controls are interesting for a company to check whether the segregation of duty is efficient, we came to the following three checks:

- Are ‘Sign’ and ‘Release’ always executed by two distinct persons?


ProM [5.0]
File Mining Analysis Conversion Exports Window Help
ProM
Analysis - Originator by Task Matrix

originator	Change Line	Create PO	GR	IR	Pay	Release	Sign
	0	0	0	0	0	1	0
	0	0	0	0	0	7	57
	0	0	0	59	74	0	0
	0	0	0	0	0	14	23
	0	0	10	0	0	10	0
	0	0	0	0	0	0	4
	51	36	53	0	0	0	0
	0	0	0	0	0	6	0
	0	0	0	0	0	10	0
	0	0	0	0	0	1	1
	48	79	86	0	0	0	107
	16	3	36	0	0	0	0
	1000	1323	665	0	0	0	173
	0	0	0	0	0	65	0
	0	0	0	0	0	0	1
	0	0	0	0	0	3	18
	0	0	0	0	0	0	3
	0	0	0	0	0	0	40
	0	2	7	0	0	0	0
	0	0	0	0	0	7	0
	30	53	53	0	0	0	52
	0	0	0	0	0	1733	1512
	0	0	0	0	0	3	0
	0	0	0	0	0	0	1
	1	0	0	0	0	0	0
	0	0	0	0	0	0	34
	94	171	58	0	0	0	0
	0	0	19	0	0	0	0
	7	6	0	0	0	0	0
	0	0	42	0	0	0	0
	0	0	0	0	0	0	2
	0	0	0	0	0	0	2
	0	0	2	0	0	10	9
	21	62	51	0	0	0	0
	11	8	12	0	0	0	0

13:44:53 [T] </OriginatorByTaskMatrix>

Figure 8.5: Print screen of the Originator-Task matrix

- Are ‘GR’ and ‘IR’ always executed by two distinct persons?
- Are ‘Release’ and ‘GR’ always executed by two distinct persons?

When designing the right formula to execute the first check, it is important to take into account that this has to be checked pairwise. If a release takes place, then a ‘Change Line’ occurs, and the next sign is performed by the previous releaser, this does not have to pose a problem. As long as the release, following the last sign, is given by another employee, the segregation of duty is intact. It turned out that in all cases this first proposition holds.

To examine the second proposition we checked whether there exist cases where the activities ‘GR’ and ‘IR’ are executed by the same originator. The *LTL Checker* revealed that in all cases, these activities were performed by different originators.

We used the same formula to check whether there exists a case in which ‘Release’ and ‘GR’ are performed by the same originator. This check revealed that in 21 cases the originator of ‘Release’ and ‘GR’ were the same. Running some extra analyzes on these cases, revealed that only four persons were responsible for the 21 inconsistencies (as opposed to 17 originators whom were involved in the complete audit trails of the 21 cases). One person performed ten times both activities ’Release’ and ’GR’, another person nine times, and two other persons each performed the two activities (in one case) once. The 21 cases belong only to five purchasing groups.

Where in the process diagnostic step the role analysis is restricted to a role-activity matrix, it is interesting to have a closer look at the collaboration between the 17 originators involved, because this is a smaller, manageable group. Having a look at the handover of work with the *Social Network Miner*, reveals the picture in Figure 8.6. The true user ID’s of the originators are erased for reasons of confidentiality. Hence the empty rectangles present 17 unique originators. It is clear there are four groups of collaboration with one person belonging to two of them. These relationships can be looked into by the domain expert to test whether they match the designed procedures.

8.2.2 Case Specific Checks

Also some very specific checks, related to the company under investigation, can be formulated. For Epsilon for example, there is always a ’Sign’ needed before a ’Release’ can be given, except in two situations:
- The PO document type has a certain ‘value A’
and the total PO value is less than ‘amount B’.
- The supplier is ‘X’ and the total PO value is less than ‘amount C’.

Because we already found evidence of cases where no sign occurred, we checked these properties using the *LTL Checker*. At first, we got 938 incorrect cases, but this was due to an error in ProM regarding the handling of floating points. After changing the amounts into cents, ProM had no problem in checking these constraints. Nevertheless, 259 cases were found not to follow this rule. This is over 2.5% of the random sample. There were 34 originators involved. How Epsilon will deal with this bigger anomaly is not known yet.

At this case study we only formulated one case specific check. This can of course be elaborated, depending on the case company.

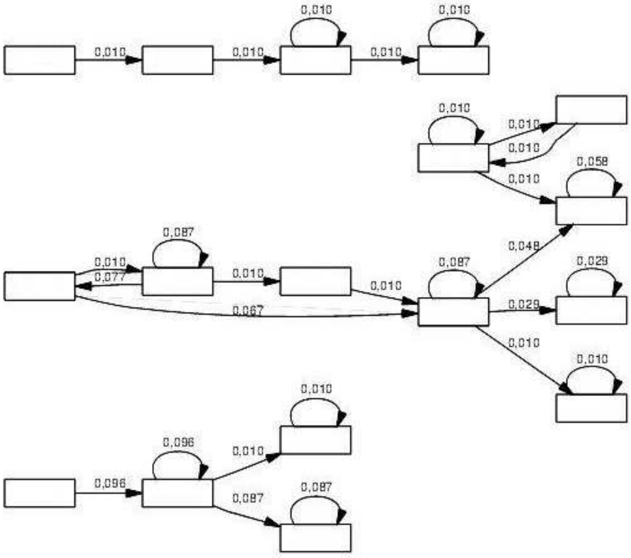


Figure 8.6: Result of *Social Network Miner - Handover of Work* for cases where 'Release' and 'GR' were not separated

8.2.3 Other Internal Control Checks

In this case study, we selected four remaining internal controls to check. The first internal control we wish to test for is very straightforward: Is every case in this event log released at least once? This is the minimum of authorization each process instance must have passed. The second check tests whether it is ensured that no payment can occur without having a corresponding invoice entered into the system. The third internal control checks tests whether the Goods Receipt Indicator is indeed turned off when no 'GR' is found in an audit trail. The fourth check tests whether the change of a PO line item appropriately triggers a new sign and/or release.

The first control of at least one 'Release' per process instance is not inspired by the process diagnostic step. However, this is a fundamental step in the process design, which makes it vital to the verification of internal controls. Running this first check reveals one case that has no 'Sign' and no 'Release' activity. The process instance was started by a batch file, but nevertheless an invoice was entered and paid afterwards, without any authorization at all. Later, a credit note was entered for the same amount. Epsilon needs to examine how this case got through the system.

For the second control, is there for each 'Pay' a corresponding 'IR'?, we have to use the attributes 'Reference Pay' and 'Reference IR' of the events 'IR' and 'Pay' respectively. Running the appropriate formula gave us first 14 cases with in total 71 payments that did not have a corresponding invoice. When looking closer at these payments, there were a lot of them created by a batch file. Eliminating these payments resulted in ten cases (encompassing 32 stand alone payments). Studying the seven involved originators (only the originators of the 'Pay' activities) revealed that one person was responsible for 22 stand alone payments. This makes it less time consuming to examine all 32 payments. It could be that in all these cases a different document than an Invoice Receipt is used, such as the Subsequent Debit, which is not included in our data base. Only manual examination can rule this out. After manual examination, it appeared all payments were based on a Subsequent Debit.

For the third check, is the Goods Receipt Indicator turned off when no 'GR' is found?, we write a corresponding LTL formula. Of the 10,000 cases, one turned out not to respect this rule.

The fourth check, is a ’Change Line’ appropriately followed by a ’Sign - Release’?, refers to the following rules: if an order is about 12,500 euros or less, an order change up to 5% is permitted to be made by an employee without approval. If an order has a total value between 12,500 and 125,000 euros, there is only a 2% freedom in modification. An order above 125,000 euros can’t be modified without approval. Above these ranges, a new authorization procedure is started. This was explained before in Chapter 3. For this check we use the attributes ’Net Value’ of the process instance and the ’Relative Modification’ and ’Modification’ attributes of the activity ’Change Line’. 77 cases turned out not to respect the above described rules. This is a low number of cases relative to 10,000 (only 0.77%).

8.3 Process Diagnostics January

As mentioned in the introduction, the verification step would be extremely interesting if it is applicable to complete data sets instead of random samples. This way a company can run for example its internal control checks monthly and have the assurance that all invoices meet all the supposed conditions. This is why we test the applicability of this idea on the data of one month, January 2007. Although the stress lies on the verification step, we still start with the first phases of process diagnostics. Phase four and five of process diagnostics (performance analysis and role analysis) are not conducted, because there is no added value in performing them again on this data set.

8.3.1 Log Preparation

All process instances that started in January (Timestamp of ’Create PO’ in January) are selected. The cleaning step of cutting all activities after the last ’Pay’ is performed.

8.3.2 Log Inspection

There are 26,185 cases in the log, containing 181,845 events. The log consists of the same seven event classes as the random sample and 272 originators are involved. A case has a minimum of four events, a maximum of 390 events, and on average six events. Table 8.5 summarizes the occurrences of the seven event classes. The same

Table 8.5: Log events January

WFMElt	Occurrences (absolute)	Occurrences (relative)
Pay	31,817	17.497%
IR	29,255	16.088%
Release	28,748	15.809%
Create PO	26,185	14.4%
Sign	25,648	14.104%
GR	24,724	13.596%
Change Line	15,468	8.506%

order of relative occurrences is found as in the random sample.

For a first inspection of the general process followed in January, we run again the *FuzzyMiner*, as we did with the random sample. The result is depicted in Figure 8.7 and reveals *Create PO-Sign-Release-GR-IR-Pay* as the most frequent path, with a side path to *Change Line*. This corresponds with the designed model, and only differs from the random sample (general) model in that it includes 'GR' in the core process (as opposed to a side path).

8.3.3 Control Flow Analysis

Uncovering the core process

Having a look at the *Performance Sequence Analysis* result, gives us some further insight in the comparison of the data set January with the random sample. In this set there are 304 patterns. This is almost twice as much as the random sample of 2007 (i.e. 170 patterns). The most frequent patterns are displayed in Table 8.6. Pattern 6 differs from pattern 0 in that the 'IR' and 'Pay' activity occur simultaneously in pattern 6, whereas the 'Pay' activity in pattern 0 takes more time. We again find a small number of patterns, three out of 304, to represent 80% of the data set.

Comparing this Table 8.6 with Table 8.2 shows that all but one frequent patterns of January also occurred frequently in the random sample. Only *Create PO-Change Line-Sign-Release-IR-GR-Pay* does not occur exactly the same in the top patterns of the random sample. However, leaving out the 'Change Line' activity, gives us

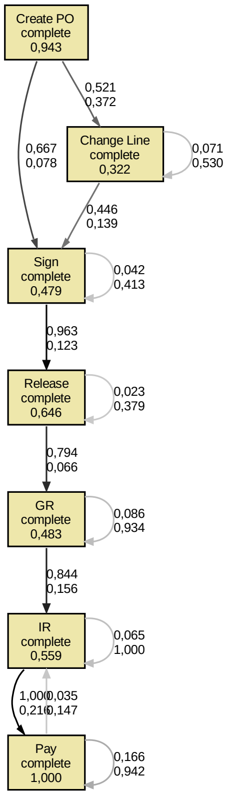


Figure 8.7: *FuzzyMiner* result January with default settings

pattern 5 of the random sample. The main characteristic of both patterns (with or without 'Change Line') remains the order of *IR-GR* instead of the other way around. So it is safe to say that all frequent patterns of the January set are also frequently present in the total data set (assuming the random sample is a good representation). The other way around, there are two patterns frequently present in the random sample, but not in the January data set: *Create PO-Sign-Release-IR-Pay* and *Create PO-Sign-Release-Change Line-IR-Pay*. The most important characteristic of the former is the absence of 'GR' and of the latter is the 'Change Line' in the middle of the process. Where the random sample showed a regular pattern without 'GR' as most occurring (*Create PO-Sign-Release-IR-Pay*, 30.7%), January does not show this pattern even in its top 7. This does not come as a surprise, there also the general model, depicted in Figure 8.7, showed 'GR' as an activity of the core process model, and not merely as an 'option' as with the random sample. We can conclude this is the difference most present between the random sample and the January data set. The fact that pattern 6 of the random sample (*Create PO-Sign-Release-Change Line-IR-Pay*) is not represented in January's top patterns, is not of that much importance.

Table 8.6: Top 7 of most occurring sequences January

		Occurrences		Total
Pattern	Sequence	#	%	%
0	*Create PO - Sign - Release - GR - IR - Pay*	11,608	44.3%	44%
1	*Create PO - Change Line - Sign - Release - GR - IR - Pay*	6,955	26.6%	71%
2	*Create PO - Change Line - Release - IR - Pay*	2,488	9.5%	80%
3	*Create PO - Release - IR - Pay*	640	2.4%	83%
4	*Create PO - Change Line - Sign - Release - IR - Pay*	491	1.9%	85%
5	*Create PO - Change Line - Sign - Release - IR - GR - Pay*	393	1.5%	86%
6	*Create PO - Sign - Release - GR - IR - Pay*	236	0.9%	87%

Running a conformance check on the January data is not interesting. For the purpose of conformance checking it is better to use a more general data set (like the random sample of a complete year) than the data of one month. As a consequence, this step is not performed on this data. On the other hand, it is interesting to have a look at the extra flows the January set reveals when lowering the *FuzzyMiner* thresholds. This again can give us information on the need to turn to the verification step or not. In the following paragraph we lower the *FuzzyMiner* thresholds to expose less frequent flows.

Exposing less frequent flows

Like with the random sample, we set the 'Cutoff' edge filter to 0.85. The result is depicted in Figure 8.8. Seven extra flows appear, which are immediately checked with the *LTL-Checker* if they indeed occur directly in this order. The flows with the corresponding results of the *LTL-Checker* (number of occurrences and consequence) are given in Table 8.7. Only 'Create PO → GR' does not explicitly occur in the data set, the other flows do. 'Create PO → Release' and 'Change Line → Release' show the need to conduct the case specific check. In this check, specific cases where 'Sign' can be skipped, are controlled. The 11 cases where 'Sign → GR' occurs are related to five PO's. These PO's have to be manually checked by Epsilon. A first quick view however was reassuring in that a 'Release' did take place, only at another moment. The flows 'Release → IR' and 'Release → Pay' both stress the importance of the Goods Receipt Indicator check; and the flows 'Release → Pay' (again) and 'Pay → IR' stress the added value of the internal control check whether for each 'Pay' there exists a corresponding 'IR'. The Role Analysis, not described in this section for the January data set, showed again the urge for a check on the segregation of duty. All these checks together bring us to the verification step, performed in the next section.

8.4 Verification Step January

We take exactly the same checks as with the random sample. The purpose is to look whether it is a realistic scenario for a company in terms of computability and in terms of manageable outcome to run such checks each month.

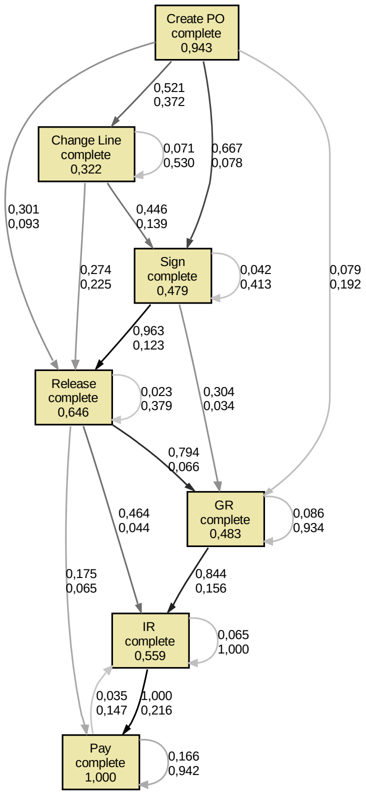


Figure 8.8: *FuzzyMiner* result January with 'Cutoff'=0.85

Table 8.7: Extra flows in January with *FuzzyMiner* 'Cutoff'=0.85

Extra flow	Occurrences	Result
Create PO → GR	0	OK
Create PO → Release	739	input Verification
Change Line → Release	2,790	input Verification
Sign → GR	11	manual check
Release → IR	4,973	input Verification
Release → Pay	244	input Verification
Pay → IR	227	input Verification

8.4.1 Checks on Segregation of Duty

The checks whether 'Sign' and 'Release', and 'GR' and 'IR' always are executed by different originators, showed these conditions were respected in all cases. This was the same result as at the random sample.

Whether 'Release' and 'GR' always are executed by different originators is also checked. Where in the random sample only 21 cases were found, not complying with the rules, there were found 175 cases in the January set, with 24 originators involved. At first sight this looks time consuming to investigate manually, but the 175 cases can be brought back to three persons. One of them performed 129 times both activities, a second person 42 times, and a last person 4 times.

As with the random sample, also here it is interesting to have a closer look at the collaboration between these originators. Having a look at the handover of work between the 24 employees with the *Social Network Miner*, reveals the picture in Figure 8.9. The result shows three groups of collaboration and one person working alone. These relationships can again be looked into by the domain expert to test whether they match the designed procedures.

8.4.2 Case Specific Check

The case specific check concentrates on the process instances where no 'Sign' was found. It is possible that PO's are not signed before they are released, but then there are some conditions which should be met. This check examines whether these conditions are met. Encouraged by the extra flows 'Create PO → Release' and 'Change Line → Release' exposed in a former step, we conduct this case specific check. Again, as with the random sample, this check revealed a larger group than manually feasible to check: 742 cases were classified as *incorrect*. This is 2.8% of the total January data set, comparable with the 2.5% found at the random sample. Comparing the originators involved (34 originators of the random sample and 30 originators from January), there is a large overlap. Only two originators involved in the sub sample of the random sample and also present in the complete January set, are not present in the sub sample of January. Because not all originators involved in the 'violating' cases are indeed violating some rules, it is very logic that there is no complete overlap. Because both data sets reveal a slightly larger group of incorrect instances, a discussion with the domain expert about the conditions of 'no Sign necessary' is recommended.

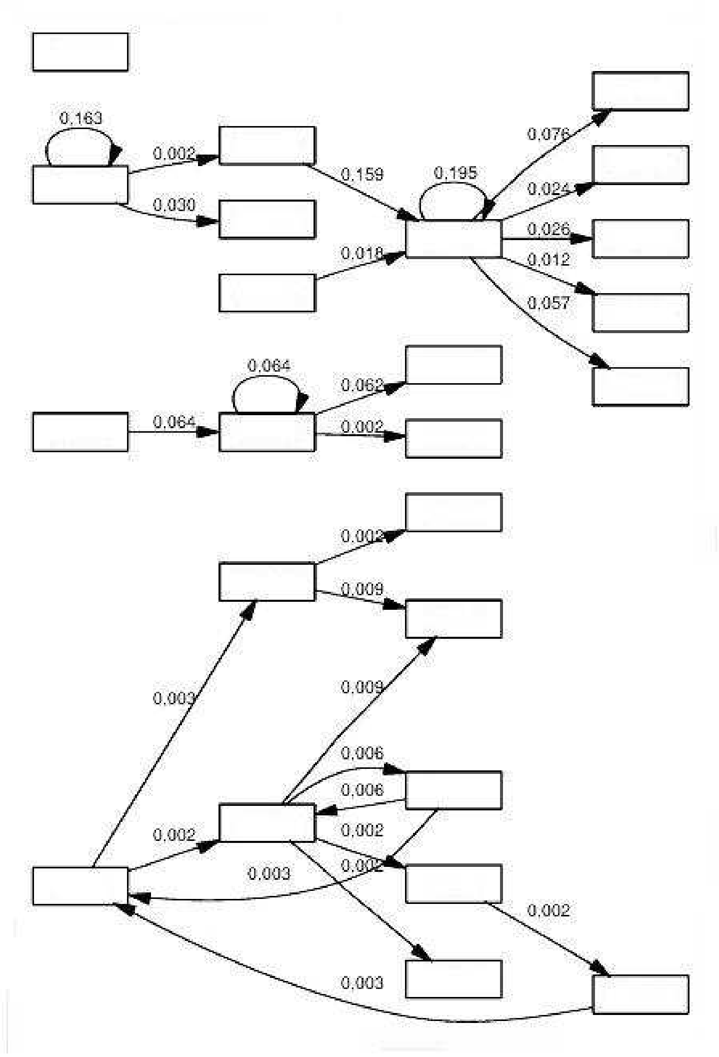


Figure 8.9: Result of *Social Network Miner - Handover of Work* for January where 'Release' and 'GR' were not separated

Possibly there are other exceptions that were not taken into account in the check. After discussing these possible extra conditions, the check can be adapted.

8.4.3 Other Internal Control Checks

As a start, each case is searched on at least one 'Release' activity. Out of the 26,185 cases, two were found not to have any authorization (no 'Release' and no 'Sign'). They showed the same pattern as the one in the random sample: created by a batch file and later paid and reversed by other originators.

The next internal control we check is whether there exists an appropriate invoice for each payment. Again, this check involved a cleaning step in filtering out the batch originators. This resulted in 46 incorrect process instances, encompassing 265 stand alone payments. One process instance has 131 'Pay' activities without a corresponding 'IR', another 75, and yet another 10. The remaining process instances only have one, two or three stand alone payments. There were 17 originators responsible for these payments. One of these originators is responsible for 216 out of the 265 payments. Two other originators have respectively 18 and 12 stand alone payments on their account. Again, all these payments are sorted out manually to check whether the payments were based on a Subsequent Debit. This seemed to be the case with all payments.

The third internal control we wish to test for is whether the Goods Receipt Indicator is turned off whenever a PO shows no GR. At the random sample there was one case that did not comply with this condition. In the January set, there are three cases found.

The last internal control setting we test for is whether a 'Change Line' triggers an appropriate Sign/Release reaction, depending on the magnitude of the change. This analysis revealed 448 incorrect cases, representing 1.71% of the January data set.

8.5 Conclusion

In this chapter we apply the process mining part of the Extended IFR^2 Framework at Epsilon. We started with a process diagnostic step of a random sample of 2007 and turned afterwards to a verification step of this sample. At this verification step, we performed tests on the segregation of duty, a case specific check, and some other internal controls were monitored. For this random sample very interesting results came to light. Only two checks revealed a higher number of cases, which may be too time consuming to examine completely manually. This decision is however Epsilon's to make. Further were both the diagnostic step and the verification step helpful in shedding a light on the procurement process and its possible weaknesses. Less frequently followed flows showed that extra tests are indeed necessary to monitor the outcome of the system, instead of fully relying on the ERP system.

Because the verification step can be of tremendous value for an organization to perform on, for instance, a monthly basis, an application test was run for the data of January 2007. Again the same two tests (the case specific check and whether a 'Change Line' triggers the right action) revealed substantively more cases than the other tests and should be discussed with the domain expert and updated in order to fine tune these checks. Altogether, both the reasonable computation time and the relatively manageable outcome (all but two tests) are in favor of a timely execution of the verification step, hereby following a more profound approach to internal control monitoring.

Chapter 9

Discussion

This last chapter reflects on the work executed in this thesis. First, a summary of the normative research part of this thesis, is given in Section 9.1. This is followed by a summary of the empirical research conducted at Epsilon in Section 9.2. In Section 9.3, criticisms and suggestions for further research are given, followed by the implications for theory and practice in Section 9.4. In a last section, some concluding thoughts are given, inspired by the results of this thesis.

9.1 Summary Normative Research

There is nothing new about saying that fraud is an important part of business. It is a multi-million dollar business concern, as several research studies reveal and as reflected in recent surveys by the Association of Certified Fraud Examiners (ACFE, 2008) and PriceWaterhouse&Coopers (PwC, 2007). Triggered by the never ending flow of scandals at the news, we decided to face the challenge as academics to contribute in the fight against internal fraud by prevention and detection. Starting from an accounting field, we turned to the information systems field to enrich our domain. After consulting current academic literature, it soon became clear this subject was not investigated very thoroughly yet. External fraud however was. By combining the academic research on external fraud and business practice lessons concerning internal fraud, we established a methodology for our research. This was poured into a framework for internal fraud risk reduction, called the IFR^2 Framework. The main idea of this framework is to introduce the field of data mining into the fight against internal fraud.

After applying the IFR^2 Framework at a case company, we turned back to the framework to extend this with another research field: process mining. Process mining aims at uncovering *a posteriori* a process model, based on an event log from the business process under consideration. Several advantages in the context of internal fraud risk reduction were identified in this field, and hence this was added as an extra part of our IFR^2 Framework.

The Extended IFR^2 Framework for internal fraud risk reduction holds a methodology to conduct research in this field. An extended literature review of current academic work was conducted. A literature gap concerning internal fraud and concerning fraud prevention was accordingly exposed. Business practice on the other hand is more intensively occupied by these issues. In a next step, the results of practitioners' fight against internal fraud were studied and discussed. Afterwards, the IFR^2 Framework was deduced from both academic literature, however focused on external fraud, and from current practice in mitigating internal fraud. The academic literature provided proof of the additional value data mining techniques can offer in the field of fraud detection. Current practice revealed a framework of internal control, applied by business practice, to prevent internal fraud. Taking into account the differences between the academic literature and our purpose (internal fraud detection and prevention), the IFR^2 Framework was developed as a complement to the existing internal control framework of the COSO. The IFR^2 Framework primarily suggests to apply a descriptive data mining approach as a complement to internal control.

Later on, this framework was extended by introducing the field of process mining, leading to the Extended IFR^2 Framework. Process mining aims at extracting *a posteriori* a process model from an event log. This is in contrast to the process model in the design phase of a business process. The process insights obtained by this method of mining, are valuable in the light of fraud prevention. Having a clearer view on what actually happens, can reveal undesired practices a company was not aware of or can draw attention to inefficient procedures. This on its turn can lead to reprimanding involved employees or re-engineering some procedures, all in the light of preventing fraud. Process mining also provides us with the possibility to test certain assertions case by case. For instance, process mining makes it possible to test whether the segregation of duty principles hold for each case separately. Finding cases where an assertion, against expectations, does not hold, can be part of fraud detection. Accordingly, process mining is added to the IFR^2 Framework. The Extended IFR^2

Framework is depicted in Figure 9.1.

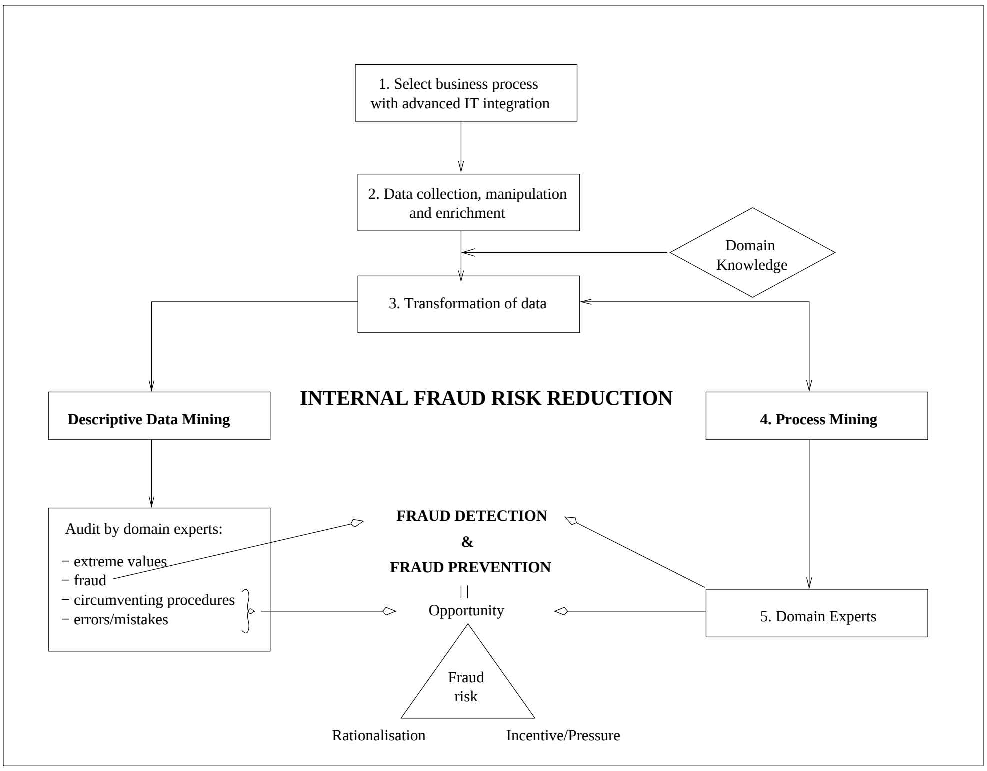


Figure 9.1: The Extended IFR^2 Framework

The Extended IFR^2 Framework is a contribution to the academic literature because of several aspects. First, it examines internal fraud, a topic that has not been addressed too often in current research.[1] Second, the focus on fraud detection, currently present in the literature, is widened to the combination of detection and prevention. Companies' risk exposure would be substantially greater if they only focused on fraud detection, a reactive working method. Companies use a combination of detection and prevention controls to help minimize their fraud risk. Hence, our study provides a more comprehensive view of the real world. Third, prior research used predictive data mining or more precisely predictive classification techniques. The

[1] Hereby meaning research where quantitative data is analyzed in order to mitigate internal fraud.

purpose of these techniques is to classify whether an observation is fraudulent or not. Because we are focusing on risk reduction rather than detection, descriptive data mining is more suited. Descriptive data mining provides us with insights on the complete data set rather than only one aspect of it, i.e., fraudulent or not. This characteristic is valuable for assessing the fraud risk in selected business processes. The additional suggestion of applying the field of process mining in the context of internal fraud risk reduction is completely new to both academic literature and practice.

After constructing our methodology framework, we were able to set up a collaboration with a case company, Epsilon. At this company, we were given the freedom to successfully apply both mining legs of our Framework. In Section 9.2, the results of both the data mining and the process mining application are summarized.

9.2 Summary Results Epsilon

We had the opportunity to apply the IFR^2 Framework at a case company, Epsilon. Epsilon, that wishes to stay anonymous in this study, is an international financial service provider. Based on a process analysis, the procurement cycle was selected as business process under investigation, more precisely the creation and payment of purchasing orders (PO's). While Epsilon's procurement is not part of its core activity, this business process encompasses a large amount of employees, procedures, and money, making it interesting as subject of investigation. Further, the lack of fraud files and the high grade of automatization at the procurement departments are in favor of investigating this business process.

9.2.1 Data Mining Case Study

All PO's that resulted in an accompanying payment in 2006[2] were subject of investigation, both in the data mining and in the process mining part. First, a process analysis of the procurement process was conducted. This was done by interviewing employees (both managers and executives), consulting procurement guide lines and existing audit reports. The result was a clear overview of how the procurement departments function. Aside from this, the weak spots in the procurement process were exposed, amongst other by discussing the process with risk managers. Starting with this background information, the data mining case study was launched. The study

[2] The creation of the PO itself however may have happened before 2006.

took place in two parts: a subset of old PO's and a subset of new PO's were analyzed, with the latter subset containing the bulk of PO's. A latent class clustering technique was used as a descriptive data mining approach. For both subsets, a three clusters model was found as best describing the data. The clustering variables chosen were the number of changes conducted on a PO, the number of changes after the last release, the percentage of the latter changes that is price related and the number of price changes over a certain limit. For both subsets, one cluster out of the three only contained 1 or 2% of the complete subset. Looking at the mean values for the clustering variables in each cluster, the small cluster two times had a very fraud risky profile. The last step of our framework requests an audit of these outlier cases to classify them as an extreme value, fraud, circumvention of procedure, or a mistake. Because the small cluster of the large subset of new PO's was too large to examine manually, only the small cluster of the old PO's was audited. Closer examination by domain experts of this small cluster revealed nine circumventions of normal procedures and one error that staid unnoticed up till now (summarized in Table 9.1).

Table 9.1: Summary of investigation by domain experts.

Category	Number of cases
Extreme values	0
Fraud	0
Circumventing procedures	9
Errors/Mistakes	1

In order to confirm the need for a multivariate approach as opposed to a univariate approach, the result of the multivariate data mining approach was compared with several univariate analyzes. No other analysis was able to provide the same result as the multivariate approach.

9.2.2 Process Mining Case Study

After the data mining case study, the process mining part of the Extended IFR^2 Framework was also applied at Epsilon. The process mining approach contained two steps: a process diagnostic step and a verification step. In the former step the general process is analyzed. In the latter step, specific internal control assertions were tested.

This process mining approach is performed on both a random sample of 2007 and the complete data set of January 2007. Both studies are summarized beneath.

Process mining random sample

The random sample of 2007 contained 10,000 cases. As mentioned before, we conducted a process diagnostic step and a verification step. The process diagnostic step confirmed the general model in place at Epsilon. This means that, in general, the procurement business process follows the designed model. However, there are still a lot of different patterns present, reflecting the complexity of this process in reality. Running some analysis checks, the need for verifying certain conditions came to light. After setting these checks into place, the results were very interesting.

The checks on segregation of duty revealed that all cases respect the conditions that a combined ’Sign’ and ’Release’ are performed by different persons, as are the activities ’GR’ and ’IR’. The condition of separating the ’Release’ and ’GR’ activities is violated 21 times, which is a very small number, easily to manually examine further.

The case specific check on the other hand yielded 259 incorrect cases, bearing a larger anomaly in the system. Although this number may be perceived as a high number, it represents only 2.6% of the random sample analyzed. Epsilon still has to answer on how to respond to this situation. It could be the case that the check has to be fine tuned.

The four other internal controls that are monitored, also revealed interesting information. A first check exposed one process instance that passed the system without any approval at all. In the end, this payment was reversed by a credit note, but nevertheless this case was paid first without a sign, nor a release. The second assertion that was tested showed ten cases with 32 payments in total without an accompanying invoice. Given the absence of other documents than ’Invoice Receipt’ in our event log, these payments could all be according procedure, using another document than ’Invoice Receipt’. Manual examination made clear all payments were based on another document type indeed. The third internal control check revealed one case where no Goods Receipt was received, although the Goods Receipt Indicator was turned on. The last assertion we tested, was whether a change of a process instance is always followed by the according authorization procedure. 77 cases did not follow the rules

Table 9.2: Analysis random sample

Less frequent flows

Create PO → Release	258	input Verification
GR → Change Line	OK	
Create PO → GR	0	OK
Sign → GR	3	manual check → OK
Sign → IR	0	OK
Release → Pay	55	input Verification

Role Analysis

Input Verification - Segregation of Duty

Verification Step

Segregation of Duty		
Originator(Sign)≠Originator(Release)	OK	
Originator(GR)≠Originator(IR)	OK	
Originator(Release)≠Originator(GR)	21	4 persons

Case Specific Check		
259 cases incorrect	259	2.6%

Other Internal Control Checks		
At least one 'Release'	1	
IR → Pay	10	manual check → 0
If no GR, then GR Ind turned off	1	
Change Line → Sign/Release	77	

as discussed with the domain expert. These cases are again handed over to Epsilon to examine. Table 9.2 summarizes the results of analyzing the random sample in the context of internal fraud risk reduction.

Process mining January

The verification step of our approach (testing assertions about internal control) would be extremely valuable if a company could perform this for instance on monthly data instead of on a random sample. Accordingly, we performed the process mining analysis also on the data from January 2007 to test the applicability. We started with a short process diagnostic step, but the emphasis lies on the verification step. The analysis results are summarized in Table 9.3 and are comparable to the results of the random sample.

The checks on segregation of duty revealed again that only the segregation of 'Release' and 'GR' is not respected. This time 175 violations were found, but only three persons were involved.

The case specific check again revealed with 2.8% of the cases a relatively high number of inconsistencies. This ratio is comparable to the 2.6% at the random sample.

The four remaining internal control checks also delivered comparable results with the random sample: (1) two cases without any approval, (2) no cases without an accompanying invoice (after manual examination), (3) three cases where no Goods Receipt was received, while this was required by procedures, and (4) a relatively larger subset of 448 cases where a change was not followed by the according authorization procedure.

About the feasibility of performing this kind of analysis each month, we can conclude that this is feasible, once the creation of the event log and the checks are programmed. However, one should preserve around one day for the creation of the event log and one day computer run time for the conversion to the appropriate MXML-format. These are off course estimates for event logs with a comparable size as ours.[3] The computability speed of the checks is less troublesome. Only one test had a run

[3] Our event log contained 26,185 process instances and 29 additional attributes

time of several hours, all other tests were run within a minute.

Table 9.3: Analysis January

Less frequent flows

Create PO → GR	0	OK
Create PO → Release	739	input Verification
Change Line → Release	2,790	input Verification
Sign → GR	11	manual check
Release → IR	4,973	input Verification
Release → Pay	244	input Verification
Pay → IR	227	input Verification

Role Analysis

Input Verification - Segregation of Duty

Verification Step

Segregation of Duty		
Originator(Sign)≠Originator(Release)	OK	
Originator(GR)≠Originator(IR)	OK	
Originator(Release)≠Originator(GR)	175	3 persons

Case Specific Check		
742 cases incorrect	742	2.8%

Internal Control Checks		
At least one 'Release'	3	
IR → Pay	46	manual check → 0
If no GR, then GR Ind turned off	3	
Change Line → Sign/Release	448	

9.3 Critical Remarks and Suggestions for Further Research

Like every research study, this dissertation is subject to several limitations. The data mining and process mining case studies are both performed on one case company. This brings along typical limitations, associated with case studies. The most important limitation is the lack of generalization. Based on case studies, we have to be careful with the interpretation of the results and beware of the danger of overgeneralization. This critique however has to be countered to a certain level. Because the case company's characteristics, relevant for this study, can be found in a lot of similar companies, we may generalize in a certain degree. The relevant characteristics, such as a SAP ERP system, a well organized procurement department and procurement as a supporting activity instead of a core activity, are to be found in a lot of companies. Also because of the standard procurement process design within SAP, we can carefully assume our results would be similar in companies that are similar to Epsilon on ERP system, organization of procurement and procurement activity. Of course, further research has to prove this.

Also some actions during the data collection are susceptible to comments. At the data mining case study, we created behavior describing attributes as a start for our analysis. Maybe other attributes could even better capture the behavior of the purchasing orders. The selection of describing attributes is inspired by the risk assessment that risks are associated with the changing behavior of PO's. A risk assessment however, although conducted with precision, can not cover all possible risks. Other assessed risks may have triggered the creation of other describing attributes than the ones in this study. Also, the data at hand shows limitations. for instance, the lack of authorization history is a valuable shortcoming in attributes. If authorization tables of the past were available, attributes about what kind of person approved could have been created, like for example 'approved by first approver' or 'approved by back-up'. These attributes were not possible to make.

On the creation of the event log for the process mining case study, the two most important remarks are the selection of a PO item line as process instance and the double dimensionality SAP works with. The selection of a PO item line as process instance was partly arbitrary. Other choices for process instance, such as an invoice line item or a complete invoice, could return different results. Also a different ERP system, without the double dimensionality of header and item information, could render different results. Further, other document types than the Invoice Receipt should be integrated in the event log. This would yield cleaner results on the check whether each payment is justified by an accompanying document. Now, this check needed further manual examination.

The order of the case studies is also susceptible to criticism. Although the data mining branch of the IFR^2 Framework can perfectly be conducted on its own, in combination with the process mining branch it is preferable to start with the process mining study. This way, one is able to lay bare weaknesses in the process to further investigate by the data mining approach. This is not a prerequisite however.

One type of fraud in particular stays extremely difficult to filter out, even with the suggested framework: fraud by collusion between employees or by collusion between an employee and some external party. Typically in this kind of fraud, procedures are not trespassed. That way, the fraud is far more difficult to detect or to prevent. The aspect that may still blow their cover, is the collusion between persons that are not supposed to work together, or at least not this intensively. This close collaboration could raise the surface at the organizational process mining perspective, but will otherwise stay unnoticed.

The suggestions for further research are inspired by both the results and the critical remarks of our research. Our framework could be further used as a methodology for research in internal fraud risk reduction. Also an evaluation framework could be added. Both on the data mining and the process mining branch, comparative studies concerning the techniques could be performed, leading to data mining and process mining techniques that are best suited in this context. Further, the implementation of a double loop expert system, could be investigated. The 'double loop' would take the output of the suggested expert system as input to filter the found anomalies even more. On the process mining side of the story, research towards the best suited selection of process instance is desired. Along with the research to best techniques and best selection of process instance, case studies in more and different companies are welcome, in order to generalize or maybe adapt the framework. And although this thesis focused on internal fraud risk reduction, a more narrow focus, we hope to expand this research to the field of continuous monitoring.

9.4 Implications for Theory and Practice

The studies conducted in this dissertation have their implications for both theory and business practice.

By constructing the Extended IFR^2 Framework, a framework for further research in internal fraud is provided. Both the topic of internal fraud in a data analysis context and the focus on fraud risk reduction are a contribution to the academic literature. Further, as opposed to the external fraud literature, a descriptive data mining technique is suggested. We encourage other academics to apply the same methodology in their internal fraud research and also to widen their focus from fraud detection to fraud risk reduction.

Both the data mining and the process mining leg of the IFR^2 Framework have their implications in academic research. Efforts on constructing supervised data sets could enhance the research at suitable descriptive data mining techniques. On the other hand is the stress on trespassed procedures in the mitigation of internal fraud a new concept. The importance of procedures in the light of fraud prevention has always been known, but not yet thoroughly inserted in quantitative data analyzes, and this is what we aim at (mostly) with process mining. By analyzing the process instead of the outcome of the process, the cure is placed closer to the disease. Also, the verification step of the process mining study could have a high impact on the continuous monitoring research. Because we are able to test assertions on every single case that passes the system (instead of a random sample), this approach brings along a new area of internal control and continuous monitoring.

Aside from the academic implications, also the case company Epsilon is affected by this research. As the results of the data mining and process mining study show, it seems interesting to pay attention on the procurement process. Although not part of this study, some double payments were uncovered by applying reporting techniques. On top of this, both the data mining and the process mining study revealed several deviations from normal procedures. On the other hand, a great part of the internal controls, tested with the process mining study, were functioning very good. Two other tested internal controls need to be examined further. Taken together, Epsilon should continue to conduct reporting techniques to rule out inconsistencies on a cost effective way (reporting techniques don't require as much resources as dataprocess mining do). Further, Epsilon should take both data mining and process mining initiatives on a timely manner to continue improving their internal fraud risk reduction.

Aside academic researchers and Epsilon, also practitioners can benefit from this dissertation. The IFR^2 Framework is constructed as a way to implement the fourth and fifth components of COSO. Currently, data analysis, as part of 'control activities', is only conducted by means of reporting techniques. By our work we hope to convince companies to add a data mining and process mining dimension to their current analyzing approach. It is by no means our idea that data mining should replace reporting techniques, only to combine both ways of analyzing. The results of the data mining and process mining case studies confirm the added value of applying these analyzes. An important implication of applying our framework however, is a close cooperation between the Information Systems (IS) department and the Auditing department. Auditing aims reducing fraud (amongst other things) and the IS department will more likely analyze data stemming from the information system. When both departments complement each other with their knowledge, this will give a synergetic effect, necessary to implement the IFR^2 Framework effectively. When staying on two separated islands of IS and Auditing, the fraud problem will not be mitigated. IS needs the process analysis knowledge of Audit, and Audit needs the data analyzing skills of IS.

Another party in business practice that could be affected by our research results, is the group of fraud detection software vendors. Currently, most software packages sold in a fraud detection context offer opportunities for reporting techniques. This means that although several useful tests can be constructed, they all start with a specific fraud in mind. The (semi-)automatic process of pattern discovery, inherent to data mining, is often not present. A set of descriptive data mining techniques could be added to the current tools. This implies again a close collaboration between IS and Audit. But further research has to sort out which techniques are best suited in this context.

At last, this dissertation holds an implication to the economic community as a whole. Internal fraud represents a huge cost to the economy, as was revealed in several studies. All research conducted in this field may be useful in the search for an answer on this problem called fraud. It is however an illusion to believe fraud (both internal and external) can be ruled out by finding 'the right' method. But this does not have to stop us from searching more effective methods for fraud risk reduction. According to the 2008 Report to the Nation of the ACFE, it is estimated a company looses 7% of its annual revenues to fraud. According to the 2006 Report, a report of only two years older, this estimate was still 5%. This increase of the cost of fraud to our economy will never stop, unless an ambitious search to reduce fraud risk continues. By keeping on searching for a (partly) answer to fraud, this cost number may ever stop increasing, who knows, start decreasing.

Personal Reflection

While writing this dissertation, I learned a lot about the subject of internal fraud, of analyzing data in this context and on companies reacting on these analyzes. Not everything I learned however, is suited for a core text of a thesis. Things learned, opinions formed, ideas fed... are not based on scientific findings or are not tested to show they are right or wrong. Still, I would like to grasp the opportunity to share my most important thoughts on this subject; the thoughts that, in my opinion, matter most for further research or brainstorming.

In this dissertation, an approach for internal fraud risk reduction is suggested. While creating the IFR^2 Framework, the important value of reducing fraud opportunities came to light, as was also the case during the application of the framework. Almost all results of our case study are associated with procedures that are not followed, and hereby creating windows of opportunity to commit fraud. This raises the interesting issue of how strict procedures need to be set and need to be followed, being the issue I would like to share my ideas on. In my opinion, two main courses can be followed. The first way to operate as a company, is to set very strict procedures and to enforce employees to follow those procedures by means of strict settings of the information system. This way an organization's risk management is minimized to the follow up and control of the information system. Another approach is to leave room for flexibility, and as such for deviating procedures. This flexibility comes along with extra control activities, on top of the control monitoring of the procedures.

On a personal note, I think an organization should prefer the second approach and leave margins on procedures. These margins are not only necessary to operate efficiently (the same reason why accounting standards are not black versus white), they also serve a higher goal. If a company would enforce employees to conduct processes according to strict procedures, without any authority to make decisions on their own, employees would feel like they are treated as a child. When no responsibility is given to employees, the employees probably will not feel any responsibility at all, any form of relation towards the organization, any form of loyalty to the organization. A lack of these commitments opens the door to rationalization of fraud, one of the fraud triangle elements. So actually while trying to reduce one element, another element will form a bigger threat. It will not surprise that an employee that is given no responsibility at all, abstracts the company as something 'far from bed' and sees no harm in stealing from it. We even don't have to go that far. No commitment to the organization also means no awareness of colleagues' fraud. It is practical impossible to stimulate a fraud aware environment in a company, when there is no bond between the employee and company. So in order to create a right culture and tone at the top, an organization should offer her employees some responsibility and authorization. This can in turn create a commitment between the employee and the organization, which will be the starting point of avoiding rationalization on one hand and of creating a fraud aware environment on the other hand. The backside of this responsibility, given to employees, is that extra controls on deviating procedures need to be put in place.

Bibliography

Abidogum, O. A. (2005). *Data mining, fraud detection and mobile telecommunications: Call pattern analysis with unsupervised neural networks.* Ph. D. thesis, University of the Western Cape.

ACFE (2008). 2008 ACFE Report to the nation on occupational fraud and abuse. Technical report, Association of Certified Fraud Examiners.

Albrecht, W., C. Albrecht, and C. Albrecht (2004). Fraud and corporate executives: Agency, stewardship and broken trust. *Journal of Forensic Accounting*, 109–130.

Albrecht, W. S., K. R. Howe, and M. B. Romney (1984). *Deterring Fraud: The Internal Auditor's Perspective.* Institute of Internal Auditors Research Foundation.

Alles, M., G. Brennan, A. Kogan, and M. A. Vasarhelyi (2006). Continuous monitoring of business process controls: A pilot implementation of a continuous auditing system at Siemens. *International Journal of Accounting Information Systems 7*, 137–161.

Bermúdez, L., J. Pérez, M. Ayuso, E. Gómez, and F. Vázquez (2007). A Bayesian dichotomous model with asymmetric link for fraud in insurance. *Insurance: Mathematics and Economics 42*(2).

Bologna, G. and R. Lindquist (1995). *Fraud Auditing and Forensic Accounting.* John Wiley & Sons.

Bolton, R. and D. Hand (2001). Unsupervised profiling methods for fraud detection.

Bolton, R. and D. Hand (2002). Statistical fraud detection: A review. *Statistical Science 17*(3), 235–255.

Bonchi, F., F. Giannotti, G. Mainetto, and D. Pedreschi (1999). A classification-based methodology for planning audit strategies in fraud detection. In *Proceedings of the*

Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA. ACM Press.

Boyer, M. M. (2007). Resistance (to fraud) is futile. *The Journal of Risk and Insurance 74*(2), 461–492.

Bozkaya, M., J. Gabriels, and J. M. van der Werf (2009). Process diagnostics: A method based on process mining. *Forthcoming Proceedings of International Conference on Information, Process, and Knowledge Management (eKNOW).*

Brause, R., T. Langsdorf, and M. Hepp. Neural data mining for credit card fraud detection. *Goethe Universit*
/.

Brockett, P. L., R. A. Derrig, L. L. Golden, A. Levine, and M. Alpert (2002). Fraud classification using principal component analysis of RIDITs. *The Journal of Risk and Insurance 69*(3), 341–371.

Brockett, P. L., X. Xia, and R. A. Derrig (1998). Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. *The Journal of Risk and Insurance 65*(2), 245–274.

Burge, P. and J. Shawe-Taylor (2001). An unsupervised neural network approach to profiling the behavior of mobile phone users to use in fraud detection. *Journal of Parallel and Distributed Computing 61*, 915–925.

Cahill, M., D. Lambert, J. Pinheiro, and D. Sun (2000). Detecting fraud in the real world.

Chien, C.-F. and L.-F. Chen (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with Applications 34*(1), 280–290.

Choo, F. and K. Tan (2007). An "American Dream" theory of corporate executive fraud. *Accounting Forum 31*, 203–215.

CICA/AICPA (1999). Continuous auditing. Technical report, The Canadian Institute of Chartered Accountants.

Cortes, C., D. Pregibon, and C. Volinsky (2002). Communities of interest. *Intelligent Data Analysis 6*, 211–219.

Cosserat, G. W. (2004). *Modern Auditing* (2 ed.). John Wiley & Sons, Ltd.

Cox, K., S. Eick, G. Wills, and R. J. Brachman (1997). Viual data mining: Recognizing telephone calling fraud. *Data Mining and Knowledge Discovery 1*, 225–231.

Davey, N., S. Field, , R. Frank, P. Barson, and G. McAskie (1996). The detection of fraud in mobile phone networks. *Neural Network World 6*(4), 477–484.

Davia, H. R., P. Coggins, J. Wideman, and J. Kastantin (2000). *Accountant's Guide to Fraud Detection and Control* (2 ed.). John Wiley & Sons.

Derrig, R. A. and K. M. Ostaszewski (1995). Fuzzy techniques of pattern recognition. *The Journal of Risk and Insurance 62*(3), 447–482.

Deshmukh, A. and L. Talluru (1998). A rule-based fuzzy reasoning system for assessing the risk of management fraud. *International Journal of Intelligent Systems in Accounting, Finance & Management 7*(4), 223–241.

Dorronsoro, J., F. Ginel, C. Sanchez, and C. Santa Cruz (1997, July). Neural fraud detection in credit card operations. *IEEE Transactions on Neural Networks 8*(4), 827–834.

Elsas, P. I. (2008). X-raying segregation of duties: Support to illuminate an enterprises's immunity to solo-fraud. *International Journal of Accounting Information Systems 9*(2), 82–93.

Ernst&Young (2007). A survey into fraud risk mitigations in 13 european countries. Technical report, Ernst&Young.

Estévez, P., C. Held, and C. Perez (2006). Subscription fraud prevention in telecommunications using fuzzy rules and neural networks. *Expert Systems with Applications 31*(2), 337–344.

Ezawa, K. J. and S. W. Norton (1996). Constructing Bayesian networks to predict uncollectible telecommunications accounts. *IEEE Expert: Intelligent Systems and Their Applications 11*(5), 45–51.

Fama, E. F. and M. C. Jensen (1983, June). Separation of ownership and control. *Journal of Law & Economics XXVI*, 301–325.

Fan, W. (2004). Systematic data selection to mine concept-drifting data streams. *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Fanning, K. and K. Cogger (1998). Neural network detection of management fraud using published financial data. *International Journal of Intelligent Systems in Accounting, Finance & Management 7*, 21–41.

Fawcett, T. and F. Provost (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery 1*(3), 291–316.

Fawcett, T. and F. Provost (1999). Activity monitoring: Noticing interesting changes in behavior. In Chaudhuri and Madigan (Eds.), *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, pp. 53–62. ACM Press.

Green, B. and J. Choi (1997, Spring). Assessing the risk of management fraud through neural network technology. *Auditing 16*(1), 14–28.

Gunther, C. and W. van der Aalst (2007). Fuzzy mining: Adaptive process simplification based on multi-perspective metrics. In *Lecture Notes in Computer Science*, Volume 4714, pp. 328–343. Springer-Verlag, Berlin.

Hagenaars, J. A. and A. L. McCutcheon (2002). *Applied Latent Class Analysis*. Cambridge University Press.

He, H., J. Wang, W. Graco, and S. Hawkins (1997). Application of neural networks to detection of medical fraud. *Expert Systems with Applications 13*(4), 329–336.

Hilas, C. S. (2009). Designing an expert system for fraud detection in a private telecommunications network. *Expert Systems with Applications 36*(2), 11559–11569.

Hilas, C. S. and P. A. Mastorocostas (2008). An application of supervised and unsupervised learning approaches to telecommunications fraud detection. *Knowledge-Based Systems 21*(7), 721–726.

Hoogs, B., T. Kiehl, C. Lacomb, and D. Senturk (2007). A genetic algorithm approach to detecting temporal patterns indiciative of financial statement fraud. *Intelligent Systems in Accounting, Finance & Management 15*, 41–56.

IIA (2005). Continuous auditing: Implications for assurance, monitoring, and risk assessment. *Information Technology Controls - Global Technology Audit Guide (GTAG)*.

Jensen, M. C. and W. H. Meckling (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics 3*, 305–360.

Juszczak, P., N. M. Adams, D. J. Hand, C. Whitrow, and D. J. Weston (2008). Off-the-peg and bespoke classifiers for fraud detection. *Computational Statistics and Data Analysis 52*(9), 4521–4532.

Kaplan, D. (2004). *The Sage Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks: Sage Publications.

Kim, H. and W. J. Kwon (2006). A multi-line insurance fraud recognition system: a government-led approach in Korea. *Risk Management and Insurance Review 9*(2), 131–147.

Kirkos, E., C. Spathis, and Y. Manolopoulos (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications 32*(4).

Knechel, W. R. (2001). *Auditing Assurance & Risk* (2 ed.). South-Western College Publishing.

Konrath, L. F. (2002). *Auditing, A Risk Analysis Approach* (5 ed.). South-Western.

Laub, J. H. (2006). Edwin H. Sutherland and the Michael-Adler report: Searching for the soul of cirminology seventy years later. *Criminology 44*(2), 235–257.

Lin, J., M. Hwang, and J. Becker (2003). A fuzzy neural network for assessing the risk of fraudulent financial reporting. *Managerial Auditing Journal 18*(8), 657–665.

Lynch, A. and M. Gomaa (2003). Understanding the potential impact of information technology on the susceptibility of organizations to fraudulent employee behaviour. *International Journal of Accounting Information Systems 4*, 295–308.

Maes, S., K. Tuyls, B. Vanschoenwinkel, and B. Manderick (2002). Credit card fraud detection using Bayesian and neural networks. *Proc. of the 1st International NAISO Congress on Neuro Fuzzy Technologies, January 6-19, 2002*.

Magidson, J. and J. K. Vermunt (2002). Latent class models for clustering: A comparison with k-means. *Canadian Journal of Marketing Research*.

Major, J. and D. Riedinger (2002). EFD: A hybrid knowledge/statistical-based system for the detection of fraud. *The Journal of Risk and Insurance 69*(3), 309–324.

Michalski, R. S., I. Bratko, and M. Kubat (1998). *Machine Learning and Data Mining - Methods and Applications*. John Wiley & Sons Ltd.

Murad, U. and G. Pinkas (1999). Unsupervised profiling for identifying superimposed fraud. *Lecure Notes in Computer Science 1704*, 251–262.

NHCAA (2008, August). http://www.nhcaa.org/.

NICB (2008, August). https://www.nicb.org/.

Pathak, J., N. Vidyarthi, and S. Summers (2003). A fuzzy-based algorithm for auditors to detect element of fraud in settled insurance claims. *Odette School of Business Administration Working Paper No. 03-9*.

Pavlou, P. A., H. Liang, and Y. Xue (2007, March). Understanding and mitigating uncertainty in online exchange relationships: a principal-agent perspective. *MIS Quarterly 31*(1), 105–136.

Phua, C., D. Alahakoon, and V. Lee (2004). Minority report in fraud detection: classification of skewed data. *SIGKDD Explorations 6*(1), 50–59.

Porter, M. E. (1985). *Competitive Advantage: Creating and Sustaining Superior Performance*. Free Press.

PwC (2007). Economic crime: people, culture and controls. the 4th biennial global economic crime survey. Technical report, PriceWaterhouse&Coopers.

Quah, J. T. and M. Sriganesh (2008). Real-time credit card fraud detection using computational intelligence. *Expert Systems with Applications 35*(4), 1721–1732.

Reisig, W. (1985). *Petri Nets: An Introduction*, Volume 4 of *Monographs in Theoretical Computer Science: An EATCS Series*. Springer-Verlag, Berlin.

Rosset, S., U. Murad, E. Neumann, Y. Idan, and G. Pinkas (1999). Discovery of fraud rules for telecommunications: Challenges and solutions. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA, pp. 409–413. ACM Press.

Sánchez, D., M. Vila, L. Cerda, and J. Serrano (2008). Association rules applied to credit card fraud detection. *Expert Systems With Applications 36*(2), 3630–3640.

Stolfo, S., W. Fan, W. Lee, A. Prodromidis, and P. K. Chan (2000). Cost-based modeling for fraud and intrusion detection: Results from the JAM project. *DARPA Information Survivability Conference & Exposition 2*, 1130–1144. IEEE Computer Press, 2000.

Sutherland, E. H., D. R. Cressey, and D. F. Luckenbill (1992). *Principles of Criminology.* General Hall.

Tan, P.-N., M. Steinbach, and V. Kumar (2006). *Introduction to data mining.* Pearson Education, Inc.

Tennyson, S. and P. Salsas-Forn (2002). Claims auditing in automobile insurance: fraud detection and deterrence objective. *The Journal of Risk and Insurance 69*(3), 289–308.

Tittle, C. R., M. J. Burke, and E. F. Jackson (1986). Modeling Sutherland's theory of differential association: Toward an empirical clarification. *Social Forces 65*(2), 405–432.

Tsung, F., Z. Zhou, and W. Jiang (2007). Applying manufacturing batch techniques to fraud detection with incomplete customer information. *IIE Transactions 39*(6), 671–680.

van der Aalst, W. and A. de Medeiros (2005). Process mining and security: Detecting anomalous process executions and checking process conformance. *Electronic Notes in Theoretical Computer Science 121*, 3–21.

van der Aalst, W., H. Rijers, A. Weijters, B. van Dongen, A. de Medeiros, M. Song, and H. Verbeek (2007, July). Business process mining: An industrial application. *Information Systems 32*(5), 712–732.

van der Aalst, W., B. van Dongen, C. Gunther, R. Mans, A. A. de Medeiros, A. Rozinat, V. Rubin, M. Song, H. Verbeek, and A. Weijters (2007). ProM 4.0: Comprehensive support for real process analysis. *Lecture Notes in Computer Science 4546*, 484–494.

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

van Dongen, B., A. de Medeiros, H. Verbeek, A. Weijters, and W. van de Aalst (2005). The ProM framework: A new era in process mining tool support. *Lecture Notes in Computer Science 3536*, 444–454.

Viaene, S., M. Ayuso, M. Guillén, D. V. Gheel, and G. Dedene (2007). Strategies for detecting fraudulent claims in the automobile insurance industry. *European Journal of Operational Research 176*, 565–583.

Viaene, S., G. Dedene, and R. Derrig (2005). Auto claim fraud detection using Bayesian learning neural networks. *Expert Systems with Applications 29*(3).

Viaene, S., R. Derrig, B. Baesens, and G. Dedene (2002). A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. *Journal of Risk and Insurance 69*(3), 373–421.

Wasserman, S. and K. Faust (1998). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Wells, J. (2005). *Principles of Fraud Examination*. John Wiley & Sons.

Whittington, O. R. and K. Pany (1998). *Principles of Auditing* (12 ed.). Irwin McGraw-Hill.

Witten, I. and E. Frank (2000). *Data mining: practical machine learning tools and techniques with Java implementations*. San Francisco, Calif.: Morgen Kaufmann.

Wolfe, D. T. and D. R. Hermanson (2004). The fraud diamond: Considering the four elements of fraud. *The CPA Journal*.

Yang, W.-S. and S.-Y. Hwang (2006). A process-mining framework for the detection of healthcare fraud and abuse. *Expert Systems with Applications 31*, 56–68.

Zaslavsky, V. and A. Strizhak (2006). Credit card fraud detection using self-organizing maps. *Information and Security 18*, 48–63.