Studies on eOrganisation and Market Engineering 4

Thomas Dreier / Rudi Studer / Christof Weinhardt (Eds.)

Information Management and Market Engineering

universitätsverlag karlsruhe

Thomas Dreier / Rudi Studer / Christof Weinhardt (Eds.)

Information Management and Market Engineering

Studies on eOrganisation and Market Engineering 4

Universität Karlsruhe (TH)

Herausgeber:

Prof. Dr. Christof Weinhardt
Prof. Dr. Thomas Dreier
Prof. Dr. Rudi Studer

Information Management and Market Engineering

Thomas Dreier
Rudi Studer
Christof Weinhardt
(Eds.)

universitätsverlag karlsruhe

Forschungsergebnisse des Graduiertenkollegs 895 „Informationswirtschaft und Market Engineering“. Gefördert mit freundlicher Unterstützung der Deutschen Forschungsgemeinschaft. Die Verantwortung für den Inhalt der Beiträge liegt bei den jeweiligen Autoren.

Impressum

Universitätsverlag Karlsruhe
c/o Universitätsbibliothek
Straße am Forum 2
D-76131 Karlsruhe
www.uvka.de

Universitätsverlag Karlsruhe 2006
Print on Demand

ISSN: 1862-8893
ISBN-13: 978-3-86644-059-3
ISBN-10: 3-86644-059-6

Preface

The papers assembled in this volume represent an overview of first research results of the Graduate School "Information Management and Market Engineering", which was established at the Universität Karlsruhe (TH) in 2004 and is financed by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).

Information Management and Market Engineering focuses on the analysis and design of electronic markets. The research takes a holistic view on the conceptualization and realization of solutions regarding the implementation, quality assurance, and further development of electronic markets as well as their integration into business processes, innovative business models, and legal frameworks. This area of research is interdisciplinary in nature, drawing from computer science, economics (both macro and micro), business engineering, operational management, and law. From an economic and business perspective, the issues investigated primarily relate to the behavior of agents in electronic markets and cover questions such as the expected outcome (e. g. allocation and prices) based on the assumed behavior of the participants and the overall impact of varying information sets. Different market institutions and business models are compared with each other, and, depending on the particular objective function, "optimal" design options are derived.

From a computer science perspective, the emphasis is on the technical realization of market platforms and on the management of knowledge. As such, system architectures, communication protocols, policy specifications and ontology engineering occupy the center of attention. The methods used in this area of market engineering originate in the fields of artificial intelligence, telematics and web services, and they facilitate the development and evaluation of modern service-oriented market architectures. Furthermore, in order to understand electronic markets in their totality, a grasp of the legal aspects involved is indispensable. On the one hand, the technical solutions and business models made possible by electronic markets must conform to the existing legal framework, particularly with respect to the rules of contracting, data protection and intellectual property. On the other hand, the legal system in turn has to react to the new information technologies and the business models enabled by electronic markets. This iterative adjustment process of technology, economic and business issues, along with the legal framework, raises many complex questions. Legal issues pertaining to fully automated contracting via intelligent software agents on electronic markets showcase the treatment of legally and technically intertwined problems.

Within the scope of the research and teaching program of the Graduate School, these different perspectives are treated holistically. The tight integration of computer science, economics and business engineering, operational management as well as legal disciplines prepares graduates to assume responsibilities as pioneers and managers in academia or practice. At the Universität Karlsruhe (TH), the Graduate School complements a special diploma course on Information Engineering and Management ("Informationswirtschaft"), established in 1997, which, in accordance with the Bologna process, is now offered as a Bachelor and Master curriculum.

The papers assembled in this book are interdisciplinary in scope and employ various methodologies. The book is therefore divided into three parts according to the major objects of Market Engineering: "Environment and Transaction Object", "Business Structure, Infrastructure, Microstructure" and "Agent Behavior and Market Outcome". The topics covered range from issues linked to electronic markets (e. g., visual ontology modelling, fraud detection, benchmark model), e-auction market places (e. g., modeling and simulating competition) and modern artificial stock markets (e. g., yield management, portfolio selection) to issues of communication, control and incentive structures (e. g., application-specific behavior and communication constraints in vehicular ad-hoc networks, entry networks, repeated decision-making, the predictive power of markets, pooling uncertainty, and a policy framework for open electronic markets). Last but not least, issues of shaping markets by means of legal instruments (e. g., trust and the law, regulating e-commerce, and promotion of open access archives) also form part of this volume, as does a paper on the re-orientation of the human personality in social networks.

Without a doubt, many people are responsible for the founding, ongoing research, and continuing success of the international PhD-Program on Information Management and Market Engineering. However, particular mention should be made of the numerous international scholars who have so far dedicated their time and best efforts at the Graduate School, amongst them Gregory Kersten, Concordia University Montreal; Lawrence Lessig, Standford University; David Messerschmidt, Electrical Engineering, Stanford University; Robert A. Schwartz, Baruch College New York; Benjamin Van Roy, Stanford University; Hal Varian, School of Information, UC Berkeley—just to name a few who gave special talks and/or lectures to the graduate students in the program during the last two years. Special thanks go to all Fellows of the Graduate School and to their supervising professors who contributed to this book. Editorial responsibilities as well as the nitty-gritty work on the manuscript were accomplished with great bravura by Steffen Lamparter and Stefan Seifert.

This book constitutes one of the first volumes of the newly founded "Series on eOrganisation and Market Engineering", which is published under the auspices of members of both the School of Economics and Business Engineering and the faculty of Computer Science and is under the editorial supervision of Professors Dreier, Studer, and Weinhardt. The series is open to all writings in this new interdisciplinary field of research. Published by the Universitätsverlag Karlsruhe, the publishing unit of the Universität Karlsruhe (TH), the series is available both in print and as a free, downloadable online version (`www.uvka.de/univerlag/volltexte/2006/147/`). It is thus disseminated via a new business model which is made possible by a particular information technology and infrastructure. In addition, the publishing contract incorporates Creative Commons licensing conditions and is thus specially adapted to the latest mode of dissemination. In sum, the series is not merely a platform for papers and research undertaken in the area of information engineering and management, but also represents first fruits of Information Management and Market Engineering.

Karlsruhe,
August 2006

Thomas Dreier
Rudi Studer
Christof Weinhardt

Contents

Part I Environment and Transaction Object

Part II Business Structure, Infrastructure, Microstructure

Part III Agent Behavior and Market Outcome

List of Contributors

Anupriya Ankolekar
Institute of Applied Informatics and
Formal Description Methods
Universität Karlsruhe (TH)
ankolekar@aifb.uni-karlsruhe.de

Christiane Barz
Institute for Economic Theory and
Operations Research
Universität Karlsruhe (TH)
barz@wior.uni-karlsruhe.de

Siegfried Berninghaus
Institute for Economic Theory and
Operations Research
Universität Karlsruhe (TH)
berninghaus@wiwi.uni-karlsruhe.de

Michael Blume
Institute of Information Systems and
Management
Universität Karlsruhe (TH)
blume@iism.uni-karlsruhe.de

Saartje Brockmans
Institute of Applied Informatics and
Formal Description Methods
Universität Karlsruhe (TH)
brockmans@aifb.uni-karlsruhe.de

Xin Chen
Institute of Information Systems and
Management
Universität Karlsruhe (TH)
chen@iism.uni-karlsruhe.de

Jörn Dermietzel
Institute of Applied Informatics and
Formal Description Methods
Universität Karlsruhe (TH)
dermietzel@aifb.uni-karlsruhe.de

Thomas Dreier
Institute of Information Law
Universität Karlsruhe (TH)
dreier@ira.uka.de

Andreas Geyer-Schulz
Institute of Information Systems and
Management
Universität Karlsruhe (TH)
geyer-schulz@iism.uni-karlsruhe.de

Yalın Gündüz
Institute for Finance, Banking and
Insurance
Universität Karlsruhe (TH)
yalin.gunduz@fbv.uni-karlsruhe.de

Hannes Hartenstein
Institute of Telematics
Universität Karlsruhe (TH)
hartenstein@tm.uni-karlsruhe.de

Pascal Hitzler
Institute of Applied Informatics and
Formal Description Methods
Universität Karlsruhe (TH)
hitzler@aifb.uni-karlsruhe.de

Moritz Killat
Institute of Telematics
Universität Karlsruhe (TH)
killat@tm.uni-karlsruhe.de

Jan Krämer
Institute for Economic Theory and Operations Research
Universität Karlsruhe (TH)
kraemer@wiwi.uni-karlsruhe.de

Jürgen Kühling
Institute of Information Law
Universität Karlsruhe (TH)
kuehling@ira.uka.de

Jonas Kunze
Institute of Information Systems and Management
Universität Karlsruhe (TH)
jonas.kunze@em.uni-karlsruhe.de

Carolina M. Laborde
Institute of Information Law
Universität Karlsruhe (TH)
laborde@ira.uka.de

Steffen Lamparter
Institute of Applied Informatics and Formal Description Methods
Universität Karlsruhe (TH)
lamparter@aifb.uni-karlsruhe.de

Ralf Löschel
Institute for Economic Theory and Operations Research
Universität Karlsruhe (TH)
loeschel@wiwi.uni-karlsruhe.de

Stefan Luckner
Institute of Information Systems and Management
Universität Karlsruhe (TH)
luckner@iism.uni-karlsruhe.de

Cora Schaefer
Institute of Information Systems and Management
Universität Karlsruhe (TH)
cora.schaefer@em.uni-karlsruhe.de

Detlef Seese
Institute of Applied Informatics and Formal Description Methods
Universität Karlsruhe (TH)
seese@aifb.uni-karlsruhe.de

Christoph Sorge
Institute of Telematics
Universität Karlsruhe (TH)
sorge@tm.uka.de

Kendra Stockmar
Institute of Information Law
Universität Karlsruhe (TH)
stockmar@ira.uka.de

Rudi Studer
Institute of Applied Informatics and Formal Description Methods
Universität Karlsruhe (TH)
studer@aifb.uni-karlsruhe.de

Thomas Stümpert
Institute of Information Systems and Management
Universität Karlsruhe (TH)
stuempert@aifb.uni-karlsruhe.de

Marliese Uhrig-Homburg
Institute for Finance, Banking and Insurance
Universität Karlsruhe (TH)
uhrig@fbv.uni-karlsruhe.de

Karl-Heinz Waldmann
Institute for Economic Theory and Operations Research
Universität Karlsruhe (TH)
waldmann@wior.uni-karlsruhe.de

Christof Weinhardt
Institute of Information Systems and Management
Universität Karlsruhe (TH)
weinhardt@iism.uni-karlsruhe.de

Martina Zitterbart
Institute of Telematics
Universität Karlsruhe (TH)
zit@tm.uka.de

Part I

Environment and Transaction Object

Electronic Commerce in Argentina: A Legal Perspective

Carolina M. Laborde

Institute of Information Law,
Universität Karlsruhe (TH)
laborde@ira.uka.de

Summary. Ever since its arrival, the Internet has spawned new issues in various branches of the law. The Internet also paved the way for a new channel for commerce: electronic commerce. Electronic commerce has proliferated rapidly and is growing steadily. New legislation must keep pace with the new technology in order for a clear framework to be established. Without a doubt, legal regulation of electronic commerce is not only necessary but critical for the ongoing development of electronic transactions. This paper covers the most relevant legislation with respect to the Internet and electronic commerce in Argentina since the advent of these technologies.

1 Introduction

Electronic commerce can be analyzed from different perspectives. In order for electronic markets to be successful, market designers need to take several elements into account. An important aspect of every institution is its legal regulation. This paper will cover the legal regulation of e-commerce in Argentina. To date, there is no legislation specifically regulating e-commerce in the nation. Does this mean that e-commerce is a field without a legal framework? Not exactly, as will be discussed below there is in fact legislation governing electronic transactions.

2 Electronic Commerce in Argentina

The development of e-commerce started in the mid-nineties in Argentina. This goes hand in hand with the relatively late appearance of the Internet in Argentina in the beginning of the nineties. According to the Argentine Chamber of Electronic Commerce (www.cace.org.ar), in 1999 there were around 1 million Internet users in Argentina; that number increased to almost 10 million by 2005 (as shown in Table[1] 1 below). Likewise, the number of transactions done electronically has steadily increased.

[1] www.cace.org.ar. Main sources: Prince & Cooke, IDC, Gartner Group, Yankee Group, Forrester, eMarketer, e-Consultingcorp, Câmara Brasilera de Comercio Electrónico and own sources of the Argentine Chamber of Electronic Commerce

Table 1: Statistical data on Internet access and electronic commerce revenue in Argentina

Statistical Data in Argentina	1999	2000	2001	2002	2003	2004	2005
Internet Users (in millions):							
Active Internet Users (in millions)	1,2	2,4	3,65	4,1	5,7	7,56	9,9
Broadband Internet access (in millions)	NS	NS	0,1	0,13	0,24	0,48	0,8
E-commerce (in billions of US$):							
B2C (in millions of US$) (US$ 1 = AR$ 2,95)	NS	23,4	79,2	240	518,7	996,6	1800
NS: Not significant							

Electronic commerce in Argentina can be characterized as a local market. Most goods and services are sold and provided within Argentine territory. Another particularity is that most of the electronic transactions are business-to-consumer (B2C). Two kinds of businesses appear in the electronic commerce scenario: first, the regular stores that went online and found a new channel for commercializing its goods and services on the Internet and second, the purely online companies which offer a wide variety of services only through the Internet. In the latter category, we find pages similar to those existing all over the world: auction sites, travel agencies, etc. Business-to-business transactions (B2B) are predominantly entered into “offline”. The Internet remains a place for advertising products and services, but it is still not the preferred means for contracting.

Open Electronic Market (Mercado Abierto Electrónico)

Open Electronic Market (Mercado Abierto Electrónico). In the Open Electronic Market public and private commercial papers such as public and private bonds, treasury bonds, and provincial and county bonds are negotiated. The Open Electronic Market aims at reducing the transaction risks, offering complete and reliable price information and registering all daily transactions.

General Resolution 121/1987 of the National Securities Commission (Official Gazette of May 27, 1988), although later derogated and superseded by General Resolution 147/1990 (Official Gazette of August 7, 1990), established the groundwork for the electronic transactions of commercial papers. General Resolution 121 established that as of December 1988, the agents of the open market would have an open access information system displaying the information relating to current offers to buy and/or sell, the sales transactions occurring during the day and the total number of operations at the end of the day. The resolution further stipulated that in order to comply with the above-mentioned obligations, a computer system could be organized so that all the agents as well as the National Securities Commission would have simultaneous access to the record of offers and transactions entered into. This computer system could be implemented by means of a sociedad anónima (corporation). Accordingly, the Mercado

Abierto Electrónico S.A. (Open Electronic Market Corporation), `www.mae.com.ar`, was created and has been in operation since 1989.

3 Electronic Commerce under Argentine Law

As stated earlier, there is no statute directly regulating e-commerce, which leaves plenty of room for discussing, planning and designing an adequate legal regime. Foreign legislation on this matter will certainly be helpful and taken into account in this process; however, legislation must always be tailored to the place in which it will be applied.

The lack of a statute specifically regulating electronic commerce does not mean that this is a field without rules. Contract law applies to contracts executed via electronic means. Moreover, several statutes governing aspects directly related to electronic commerce have been passed.

3.1 National Constitution

The study of any legal topic has as a starting point the supreme law of the land, which in Argentina´s case is the National Constitution of 1853 and its amendments. Section 42 of the National Constitution, introduced by the constitutional amendment of 1994, establishes that in matters of consumption,"consumers and users of goods and services have the right to the protection of their health, safety, and economic interests; to adequate and truthful information; to freedom of choice and equitable and reliable treatment". Protection of the consumer is therefore a paramount consideration and shall be taken into account in any statute regulating electronic commerce.

3.2 National Legislation

Internet Regulation

The first set of national rules enacted were in reference to the Internet. There is a close interrelationship between the Internet and electronic commerce as access to the internet is generally needed in order to enter into transactions online. Therefore, it is worth presenting a brief review of the main regulations concerning the Internet.

The Presidential Decree 554/1997 (published in the Official Gazette of June 23, 1997) declared Internet access for Argentine citizens to be in the national interest and established that said access should be equal across social and geographical lines and provided at reasonable tariffs. That same year Presidential Decree 1279/1997 (published in the Official Gazette of December 1, 1997) declared that the Internet was protected by the constitutional guarantee of freedom of expression.

In order to achieve the goal of promoting access to the Internet, the programme "argentin@internet.todos" was launched (Presidential Decree 1018/1998 published in the Official Gazette of September 7, 1998). Within the framework of the said programme the Community Technology Centers (known as CTC) were created. The CTC are a means of providing access to computers and to the Internet as well as courses on how to use those tools to persons, living in isolated places or lacking the financial means to otherwise have access to them.

Internet 2 was developed in the United States in the mid-nineties and links universities with corporations and government agencies. The goal of Internet 2 is to foster the development of Internet technologies. Argentina would like to follow suit and interconnect its universities and scientific and technology centers. To this end, the Resolution 999/1998 drafted by the Secretary of Communications (Official Gazette of April 13, 1998) enabled the launch of Internet 2 Argentina in order to develop a high-speed network.

Another important step towards the promotion of the Internet was the government's decision to "go online". At the end of the 1990s and the beginning of the new millennium, most of the governmental bodies launched their Web sites. Though quite simple at their inception, some of them have gone on to become useful tools for research. Some of the most useful Web sites, at least from a lawyer's perspective, include the site of the Supreme Court of Justice, `www.csjn.gov.ar`, the National Official Gazette site `www.boletinoficial.gov.ar`, and the legislative information site of the Ministry of Economy, `infoleg.mecon.gov.ar`.

Consumer Protection Act

The Consumer Protection Act (24.240) was passed in 1993 (Official Gazette of October 15, 1993). Although the Internet was in its infancy in Argentina at that time, this act nevertheless contains provisions that apply to transactions conducted on the Internet. It stipulates that in a contract for sale entered into by electronic means (in which both offer and acceptance take place electronically), the buyer has the right to revoke acceptance within five days of receipt of the goods or the execution of the contract, whichever occurs first. The right to revoke acceptance cannot be waived by the buyer, and the seller is obligated to advise the buyer of this right in writing. If the buyer opts to exercises this right, the seller must bear the return costs.

The question is when the Argentine Consumer Protection Act applies. When a consumer buys a product in a store, for example and the transaction falls within the scope of the act, the consumer is protected by the consumer law. Likewise, if a consumer buys a product on a Web site offering products to Argentine consumers, then the Argentine law also applies. However, if an Argentine consumer buys a product from a non-Argentine Web site, is that consumer protected by the Argentine Consumer Protection Act? Which statutes apply to such a transaction in terms of consumer protection? In this case, the rules of private international law apply.

Data Protection Act

The Data Protection Act (25.326, Official Gazette of November 2, 2000), protects the personal data of individuals and legal entities collected in private and public databases. The Internet has proven to be an excellent conduit for gathering enormous amounts of personal data. This information can then used for various purposes, in most cases without the owner's awareness that his or her data is even being stored in a database.

To address this concern, the Data Protection Act mandates that the prior consent of the owner of the information be obtained to collect personal data, unless some of the exceptions listed in the statute apply (for example when the data is in the public domain, or when there is a prior contractual relationship between the owner of the database and the owner of the information). Moreover, when personal data is collected, it is mandatory to disclose the finality for which the information is being collected as well as the identity and address of the individual or entity responsible for the data base. The act also stipulates a provision entitling the owner of the information to know precisely which data pertaining to his or her person is contained in the database, and, when applicable, to request that the information be corrected or deleted.

These days it is not uncommon for a company to concentrate its database in one country or for personal information to be transferred to international bodies. The Data Protection Act takes these situations into account and establishes that the transfer of data to other countries or international organizations is forbidden unless adequate protection standards are guaranteed. In other words, the recipient entity is charged with ensuring adequate protection standards; otherwise, the transfer of personal data is not allowed. The statute does not define the concept of "adequate protection standards". The act's regulatory decree, 1558/2001 (Official Gazette of December 3, 2001), empowers the National Department for the Protection of Personal Data to evaluate the protection standards of a foreign country or international organization. The decree does not establish fixed rules for the assessment of international protection standards, but sets down a balancing test. This test takes into account the nature of the data to be transferred, the finality of the transfer, the amount of time the transferred data will be kept abroad, and the general pertinent legislation in the recipient country, including any rules applicable to the recipient international organization. The rule concerning international transfer of data does not apply however when the transfer is required for international judicial cooperation, cooperation among intelligence bodies in the fight against the organized crime, terrorism and drug trafficking or transfers made according to international treaties to which Argentina is a party.

Digital Signature Act

The National Congress of Argentina passed the Digital Signature Act (25.506) on November 14, 2001; the legislation was published in the Official Gazette on December 14, 2001. The President in turn issued Regulatory Decree 2628/2002 (Official Gazette of December 20, 2002). The Act governs the validity and legal effects of digital signature and also introduces other key concepts such as electronic signature and digital document.

At the time the Digital Signature Act was passed, law regulating the use of digital signatures already existed. Presidential Decree 427/1998 (Official Gazette of April 21, 1998) established the use of the digital signature for internal acts of the Public

Administration (including official banks and financial entities). This decree, now derogated and since superseded by the Digital Signature Act, was a relevant precedent for the introduction and promotion of the digital signature. Also, three months before the Digital Signature Act was passed, the Public Administration allowed procurement contracts to be concluded in digital format and use digital signatures.

Contracting through electronic means might sound attractive. However, there are risks deterring its use. In a traditional contract, the parties usually meet face to face and sign the contract in each other's presence, with each party retaining a copy. Signing a contract via the Internet presents an entirely different scenario. The parties may never meet and most likely will not sign the contract in each other's presence. Therefore, there is no certainty that the party claiming to have signed the contract is, in fact, one and the same. In addition, there is no guarantee that the document will not be manipulated once one party has obtained the other's signature. Digital signatures are intended to minimis these risks by ensuring both the identity of the signatory and inviolability of the digital document, guaranteeing that the person claiming to be the author of a document is one and the same and that the document has not been altered since its endorsement by the signatory.

According to the Digital Signature Act, a digital signature results from the application of a mathematical protocol requiring information known only to the signing party and under its absolute control, to the digital document. A digital signature shall be simultaneously verifiable by third parties to allow the identification of the signing party and to detect any alteration of the document after its endorsement. A signature that lacks any of the above-mentioned requirements is considered to be an electronic signature only. The electronic signature is defined as the associated electronic data identifying the signing party that does not meet some of the legal requirements for digital signature.

There are important distinctions between digital and electronic signatures. First, only digital signatures can be used whenever the law requires a handwritten signature. However, there are certain exceptions to this rule. The Digital Signature Act does not apply to the following transactions: wills and testaments, acts relating to family law, acts of extreme personal character (actos personalísimos), and acts whose formality is incompatible with the use of digital signature (e.g. a public deed). Another distinction between digital and electronic signatures is that the act protects the digital signature with legal presumptions. The Digital Signature Act presumes that a digital signature belongs to the owner of the digital certificate. Likewise, if the outcome of a verification proceeding is affirmative the law presumes that the document has not been modified since the signing.

For a digital signature to be valid, the following requirements must be met: the signature must be created during the period of validity of the digital certificate, the certificate must have been issued or recognized by a licensed certifier and the signature shall have been duly verified. A digital certificate is a digital document that has been digitally signed by a certifier, which links the verification data to its holder. For a digital certificate to be valid, it must have been issued by a licensed certifier, conform to international standards and contain the minimum information specified by the statute (identification of its holder and issuer, period of validity, and data to allow unique identification).

The act contains provisions dealing with the recognition of foreign digital certificates. A certificate is deemed foreign when issued by a foreign certifier. Foreign digital certificates are eligible for recognition when either (i) meet the requirements specified by Argentine law and a reciprocity agreement is in force between Argentina and the certificate's country of origin or (ii) are recognized by a licensed certifier in Argentina, which guarantees their validity and force according to current Argentine law and the said recognition is in turn validated by the relevant enforcement authority. Those are the core provisions of the statute. However, the statute devotes most of its sections to setting up the infrastructure needed to implement digital signatures.

The "Spy" Act

In February 2004, Law 25.873 (Official Gazette of February 9, 2004) amended the Telecommunications Act (19.798; Official Gazette of August 23, 1972) by introducing three new sections to it. In brief, these new sections require the telecommunication providers to maintain records of all communications for a period of ten years, to be provided on request to the Judicial Power or the Public Ministry (an independent body whose main function is to promote justice by defending the interests of the society). Likewise, the telecommunication provider shall register the personal data of the clients and users for the same purpose. The costs of this task are to be borne by the telecommunication provider. Finally, the amended statute declares that the government is responsible for any damage caused to third parties as a result of the enforcement of the statute. The alleged goal of these new sections was to combat kidnappings, where communications and prompt action play a critical role. The executive power, according to its constitutional powers, issued the corresponding regulatory Decree 1563/2004 (Official Gazette of November 9, 2004). The act and its regulatory decree were known as the "Spy" Act.

These were roundly criticized as overstepping the bounds of the Argentine constitution. Due to the intense criticism, the Executive Power decided to suspend the enforcement of the regulatory decree. It should be noted that the decree was not revoked but merely suspended; it is still valid but its enforcement has been delayed. In June 2005, a judge struck down both the statute and the decree on the grounds of unconstitutionality. However, under Argentine law, a statute continues to be in force despite a declaration of unconstitutionality. Such a judicial decision is only valid to the parties to the concrete case.

4 Concluding Remarks

Electronic commerce continues to develop and evolve whether there is a legal regime or not. The existence of an adequate framework would, however, inspire more confidence among users and increase the number of persons contracting on line as well as the transaction volume overall. In this process, lawmakers play a fundamental role as designers of the legal environment for electronic commerce. In the case of Argentina, legislation has addressed some of the new problems attendant to entering into commercial transactions electronically.

Towards a Policy Framework for Open Electronic Markets

Steffen Lamparter[1], Anupriya Ankolekar[1], Rudi Studer[1], and Christof Weinhardt[2]

[1] Institute of Applied Informatics and Formal Description Methods (AIFB),
Universität Karlsruhe (TH)
`{lamparter,ankolekar,studer}@aifb.uni-karlsruhe.de`

[2] Information Management and Systems,
Universität Karlsruhe (TH)
`weinhardt@iw.uni-karlsruhe.de`

Summary. In recent years, electronic markets have attracted a lot of attention as mechanisms for the efficient allocation of goods and services between buyers and sellers. However, calculating suitable allocations between buyers and sellers in these markets can be very tricky, particularly if the services and goods involved are complex and described by multiple attributes. Since such trading objects typically provide various configurations with different corresponding prices, complex pricing functions and purchase preferences have to be taken into account when computing allocations. In this paper, we present a policy description framework that draws from utility theory to capture configurable products with multiple attributes. The framework thus allows a declarative description of seller pricing policies as well as buyer preferences over these configurations. We present a machine-processible representation language for the policies and a method for evaluating different trading object configurations based on these policies as components of these framework.

1 Introduction

Electronic markets are institutions allowing the exchange of goods and services between multiple participants through global communication networks, such as the Internet. According to Neumann (2004), the design of market platforms mainly involves two components: a *communication language* defining how bids (i.e. offers and requests) can be formalized and submitted to the market mechanism, and an *outcome determination* calculated by means of an allocation (i.e. who gets which service), a pricing schema and a payment component. The design of the communication language is a substantial task requiring mutual understanding between different participants in the market; it involves trading objects which are offered by multiple parties (sellers) with different attributes and under different conditions. This is particularly true for configurable goods and services, such as computers and Web-based services. Consider, for example, a route planning Web service offering the service of computing a road route between two locations. Various configurations of the service may take into account the current traffic or weather situation when computing the route. Alternatively, the service may be configured to compute the shortest or quickest route, or one that avoids small roads and so on. Naturally, each configuration may have a different price attached. Decision-making in markets with these sorts of complex services and goods generally requires that both seller pricing functions as well as buyer scoring (preference) functions be taken into account. In the remainder of this paper, we denote rules that define the relation between configurations and prices defined by a seller as *pricing policies* and rules that define how much a buyer is willing to pay for a certain configuration as *scoring policies* (or buyer preferences).

In this paper, we consider the problem of designing a policy framework that enables the expression of such pricing and scoring policies. The framework has to meet several requirements (emphasized): First, since we are dealing with various configurable products, the policy framework must be able to describe ***multi-attributive requests and offers***, i.e. requests and offers that involve multiple attributes beyond just the price, such as quality criteria. Second, both pricing and scoring policies require the ***expression of functions*** that map configurations to a pricing or a preference structure, respectively. Note that preferences have to be measured on a cardinal scale, so that one can specify both a ranking of offers and an acceptance threshold for offers that satisfy the request to a certain degree. Third, since pricing policies have to be communicated to the buyer and/or scoring policies to the seller (e.g. in a procurement auction), ***standards-based interoperability*** becomes a crucial issue. Standardized syntax and semantics are particularly essential in open markets, where participants may use highly heterogenous information formats, where buyers and sellers dynamically join or leave the market, and where products and services are highly differentiated and change frequently. Our technology of choice for meeting the interoperability requirement is ontologies. Ontologies are also powerful enough to meet the requirements "multi-attribute requests and offers" and "expression of functions" due to the underlying logic and rule mechanism. This allows us to avoid the use of additional languages and technologies and simplifies our framework without loss of functionality or expressivity.

In the following, we introduce a policy framework that meets the requirements above. The key contribution of this work is to show how quantitative preferences can be modeled within a declarative framework for inclusion in the reasoning process. The approach should be as expressive as possible while adhering to Internet standards. Before presenting the framework we review related work to determine the extent to which the requirements are already supported by existing work (Section 2). Since our approach is based on *Utility Function policies*, we briefly sketch the fundamentals of utility-based policy representation in Section 3. In section 4, we present the ontology formalisms and ontology framework and then discuss how such policies can be declaratively represented by means of ontologies and be enforced using a semantic framework in Section 5. The paper concludes with a brief outlook in Section 6.

Table 1: Languages for specifying offers and requests.

Approach	**Requirement**		
	Multi-attributive	**Functions**	**Interoperability**
EDI/EDIFACT	(✓)	(✓)	-
XML-based Languages	✓	-	(✓)
KAoS/REI	✓	-	✓
SweetDeal	✓	(✓)	(✓)
Product/Service Catalogs	-	-	✓
Bidding Languages for CA	✓	✓	-
CPML	✓	✓	(✓)
Our Approach	✓	✓	✓

2 Related Work

In this section, we present the various existing approaches in electronic markets for modeling buyer preferences and seller offerings. We then discuss the extent to which they meet the requirements specified in Section 1. Table 1 summarizes the approaches discussed in terms of which of the three requirements they support.[1]

One of the first attempts to exchange order information within electronic markets was the Electronic Data Interchange (EDI) protocol, which serializes request and offer information according to a predefined format agreed upon by both communication parties. However, these pairwise agreements were rarely based on any standards and turned out to be labor-intensive, highly domain-dependent and inflexible, and thus did not address the interoperability requirement.

More recent approaches, such as WS-Policy (Siddharth Bajaj et al., 2004), EPAL (Ashley, P. et al., 2003) and WSPL (Anderson A. et al., 2003), use XML (eXtensible Markup Language) as a domain-independent syntax to define constraints on attributes of configurable trading objects within the context of Web service agreements. However, they are not suitable for our purposes because they only allow the expression of attribute value pairs and thus cannot be used to express seller pricing or buyer scoring functions. WS-Agreement Grid Resource Allocation and Agreement Protocol Working Group (2005) is another XML-based specification that can be used to express different valuations for configurations only with discrete attributes, however. An approach for extending WS-Agreement for expressing continuous functions is presented in Sakellariou and Yarmolenko (2005). However, the meaning of XML annotations is defined in a natural language specification which is not amenable to machine interpretation and promotes ambiguous interpretation.

One way to enable machine interpretation of buyer requests and seller offers is to specify them using a machine-interpretable ontology. Such an ontology consists of a set of vocabulary terms with a well-defined semantics provided by logical axioms constraining the interpretation of the vocabulary terms. This is the approach followed

[1] Supported requirements are indicated by check marks. Parenthesizes around check marks indicate that a requirement is only partially fulfilled.

by KAoS (Uszok et al., 2004) and REI (Kagal, 2004), which are policy languages that allow the definition of multi-attributive policies with constraints on attributes. However, these approaches are limited in that they always evaluate either to true or false and thus cannot express the scoring or pricing functions required for configurable products. More expressivity with respect to functions is provided by the SweetDeal approach presented by Grosof and Poon (2003). Similar to our approach, SweetDeal features automatic reasoning based on a formal logic. However, although SweetDeal uses a standard syntax (RuleML), its semantics is not yet standardized, which obstructs interoperability. In addition, while its underlying rule language might be capable of expressing utility-based policies, SweetDeal does not provide the required policy-specific modeling primitives directly; the rules for interpreting such policies therefore have to be added manually by the user.

A separate stream of work has focussed on developing highly expressive bidding languages for describing various kinds of attribute dependencies and valuations, particularly in the context of (combinatorial) auctions (e.g., Nisan, 2000; Boutilier and Hoos, 2001). However, they assume a closed environment and therefore, even if they do use XML-based bidding languages (Bichler and Kalagnanam, 2005), they do not address interoperability issues. Many B2B scenarios use standardized product and service taxonomies, such as UN/SPSC[2], CPV[3] or the MIT Process Handbook (Malone et al., 1997). However, these taxonomies are static and require the introduction of a new subclass in the hierarchy for each new product configuration. They are therefore clearly inapplicable in our context.

Our approach draws from utility theory to express scoring and pricing functions of market participants within an ontology-based policy framework. Our policy framework is based on existing Internet standards, namely XML for serializing request and offer documents, OWL (Web Ontology Language) (W3C, 2004) and DL-Safe SWRL-rules (Semantic Web Rule Language) (Horrocks and Patel-Schneider, 2004; Motik et al., 2005) to formalize ontology axioms. The exchanged documents can thus be interpreted by standard inference engines. Furthermore, to facilitate integration between offer/request specifications that use heterogenous ontology concepts, we base our policy ontology on the DOLCE foundational ontology (Masolo et al., 2002). DOLCE provides a high degree of axiomatisation (exact definition) of the policy ontology terms, an advantage which carries over to the policy ontology we present.

3 Utility-based Policy Representation

Policies are declarative rules that guide the decision-making process by constraining the decision space, i.e. they specify which alternatives are allowed and which are not. Kephart and Walsh (2004) refer to such policies as *Goal policies*. However, when making a decision, it is often not enough to merely know which alternatives are allowed;

[2] United Nations Standard Products and Services Code (`http://www.unspsc.org`)

[3] Common Procurement Vocabulary (`http://simap.eu.int/nomen/nomenclature_standards_en.html`)

rather, it is necessary to assess which alternative is best among the options and how good it actually is on its own terms (e.g. even the best alternative might be not good enough). Therefore, we suggest combining a declarative policy approach with utility theory, which quantifies preferences by assigning cardinal valuations to each alternative. With such *Utility Function policies*, detailed distinctions in preferences can be expressed, enabling improved decision-making between conflicting policies (as compared to traditional Goal policies) by explicitly specifying appropriate trade-offs between alternatives (Kephart and Walsh, 2004).

For our policy language, we use the following simple utility model. Assume alternatives (e.g. configurable trading objects) are described by a set of attributes $X = \{X_1 \ldots X_n\}$. Attribute values x_j of an attribute X_j are either discrete, $x_j \in \{x_{j1}, \ldots, x_{jm}\}$ or continuous, $x_j \in [min_j, max_j]$. Then the cartesian product $\mathcal{O} = X_1 \times \cdots \times X_n$ defines the potential configuration space, where $o \in \mathcal{O}$ refers to a particular configuration. Based on these definitions, a preference structure is defined by the complete, transitive, and reflexive relation $\succeq$. For example, the configuration $o_1 \in \mathcal{O}$ is preferred to $o_2 \in \mathcal{O}$ if $o_1 \succeq o_2$. The preference structure can be derived from the value function $V(o)$, where the following condition holds: $\forall o_a, o_b \in \mathcal{O} : o_a \succeq o_b \Leftrightarrow V(o_a) \geq V(o_b)$. In order to calculate the valuations $V(o)$ we apply an additive utility model, where the value functions defined in Equation 1 below are applied to aggregate the valuations derived from the individual attributes $X_1, \ldots, X_n$.

$$V(x) = \sum_{j=1}^{n} \lambda_j v_j(x_j), with \sum_{j=1}^{n} \lambda_j = 1 \tag{1}$$

For the additive value function above, we assume mutual preferential independence between the attributes (Keeney and Raiffa, 1976). Under this assumption, we can easily aggregate the utility functions $v_j(x_j)$ of the individual attributes j to obtain the overall valuation of a configuration. We believe that additive value functions are valid in many real world scenarios and might provide a good approximation, even when preferential independence does not hold exactly (Russel and Norvig, 2003). The weighting factor λ_j is normalized in the range $[0, 1]$ and allows modeling the relative importance of an attribute j.

In the context of electronic markets, Utility Function policies can be used on the buyer-side to specify preferences, assess the suitability of trading objects and derive a ranking of trading objects based on these preferences. Further, they allow the exchange of preferences with sellers which might for instance be required in procurement auctions or exchanges. On the seller-side, the pricing or cost function can be expressed and communicated to the customers very efficiently with only one message.

In the next sections, we show how a utility-based approach such as this one can be declaratively modeled within a system in which both available trading objects as well as policies are stored in a knowledge base and queries are the issued to derive relevant information (e.g. prices, product rankings, etc.).

4 A Policy Description Framework for Electronic Markets

Before introducing our ontology for representing Utility Function policies in Section 5, we first discuss the underlying formalisms as well as the upper-level modules upon which the ontology is based. Section 4.1 presents the ontology formalism OWL-DL as well as SWRL and Section 4.2 introduces the general ontology framework, which is based on the DOLCE foundational ontology.

4.1 Ontology Formalism

In recent years, *ontologies* have become an important technology for knowledge sharing in distributed heterogeneous environments, particularly in the context of the *Semantic Web*[4]. As defined by Gruber (1993), an ontology is a set of logical axioms that formally defines a shared vocabulary. By committing to a common ontology, we can enable software agents to make assertions or ask queries that are understood by other agents.

In order to guarantee that these formal definitions are understood by other parties (e.g. in the Web), the underlying logic has to be standardized. The Web Ontology Language (OWL) standardized by the World Wide Web Consortium (W3C) is a first effort in this direction. OWL-DL is a decidable fragment of OWL and is based on a family of knowledge representation formalisms called *Description Logics (DL)* (Baader et al., 2003). Consequently, our notion of an ontology is a DL knowledge base expressed via RDF/XML syntax to ensure compatibility with existing World Wide Web languages. The meaning of the modeling constructs provided by OWL-DL - like concepts, relations, datatypes, individuals and data values is formally defined via a model theoretical semantics. In other words, it is defined by relating the language syntax to a model consisting of a set of objects, denoted by a domain and an interpretation function. The latter maps ontological entities to concrete entities in the domain (Horrocks et al., 2003).

In order to define the Core Policy Ontology, we require additional modeling primitives not provided by OWL-DL. For example, we have to model triangle relations between concepts. In contrast to OWL, rule languages can be used to express such relations. The Semantic Web Rule Language (SWRL) (Horrocks and Patel-Schneider, 2004) allows us to combine rule approaches with OWL. Since reasoning with knowledge bases that contain arbitrary SWRL expression usually becomes undecidable (Horrocks and Patel-Schneider, 2004), we restrict ourself to *DL-safe* rules (Motik et al., 2005). DL-safe rules keep the reasoning decidable by placing constraints on the format of the rule; each variable occurring in the rule must also occur in a non-DL-atom in the body of the rule. This means the identity of all objects referred to in the rule has to be explicitly known. To query and reason over a knowledge base containing OWL-DL as well as DL-safe SWRL axioms, we use the KAON2 inference engine[5].

[4] `http://www.w3.org/2001/sw/`

[5] available at http://kaon2.semanticweb.org/

Table 2: Upper level concepts from DOLCE, Descriptions and Situations (DnS), Ontology of Plans (OoP) and Ontology of Information Objects (OIO) that are used as a basis for modeling.

Module	Concept label	Usage
DOLCE	Endurant	Static entities such as objects or substances
	Perdurant	Dynamic entities such as events or processes
	Quality	Basic entities that can be perceived or measured
	Region	Quality space (in this work implemented as datatypes)
DnS	Description	Non-physical objects like plans, regulations, etc. defining Roles, Courses and Parameters
	Role	Descriptive entities that are played by Endurants (e.g. a customer that is played by a certain person)
	Course	Descriptive entities that sequence Perdurants (e.g. a service invocation which sequences communication activities)
	Parameter	Descriptive entities that are valued by Regions, such as the age of customer
	Situation	Concrete real world state of affairs using ground entities from DOLCE
OoP	Task	Course that sequences Activities
	Activity	Perdurant that represents a complex action
OIO	InformationObject	Entities of abstract information like the content of a book or a story

For the reader's convenience, we define DL axioms informally via UML class diagrams[6], where UML classes correspond to OWL concepts, UML associations to object properties, UML inheritances to subconcept-relations and UML attributes to OWL datatype properties (Brockmans et al., 2004). For representing rules, we rely on the standard rule syntax as done in Horrocks et al. (2004); Motik et al. (2005). In the following, SWRL rules are labelled by $R1, \ldots, Rn$.

4.2 Modeling Basis

Our policy description framework consists of several ontology modules. These modules are composed of three layers: *(i)* As a modeling basis, we rely on the domain-independent upper-level DOLCE foundational ontology (Masolo et al., 2002). By capturing typical *ontology design patterns* (e.g. location in space and time), foundational ontologies provide basic concepts and associations for the structuring and formalization of application ontologies. Furthermore, they provide precise concept definitions and a high axiomatization. Foundational ontologies facilitate thereby the conceptual integration of different languages (Gangemi et al., 2003) and thus ensure interoperability in heterogenous environments. Due to limited space, we will omit a detailed description of DOLCE here. The DOLCE concepts that are used for the alignment of our ontology are briefly introduced in Table 2. A detailed description of DOLCE and its modules

[6] The entire ontology is also available in RDF/XML serialization at `http://ontoware.org/projects/emo`

is given in Masolo et al. (2002) and Gangemi et al. (2004b). *(ii)* As a second layer, we introduce the *Core Policy Ontology*, which extends the upper-level ontology modules by introducing concepts and associations that are fundamental for formalizing policies. *(iii)* While the first two layers contain domain-independent off-the-shelf ontologies, the third layer comprises ontologies for customizing the framework to specific domains (e.g. an ontology for modeling certain products and their attributes).

In the next section, we focus on the Core Policy Ontology and show how policies and configurations are modeled based on DOLCE and the logical formalism introduced above.

5 Core Policy Ontology

In the following, an ontology for modeling policies is introduced that meets the requirements regarding expressivity and interoperability discussed above. The remainder of this section is structured as follows: first, we extend the DOLCE ground ontology by modeling the primitives required for representing functions between attribute values and their individual valuation by a user. Based on these functions, we then show how the Description & Situation ontology design pattern is applied to model product configurations and policies. Finally, we introduce interpretation rules that evaluate the configurations according to the specified policies.

5.1 Valuation Functions

As discussed in Section 3, preferences as well as pricing information are expressed via functions that map configurations to a corresponding valuation between 0 (or $-\infty$) and 1, where a valuation of $-\infty$ refers to forbidden alternatives and a valuation of 1 to the optimal alternative (Lamparter et al., 2005). We now illustrate how the fundamental concepts formalized in DOLCE can be extended to allow the expression of valuation functions.

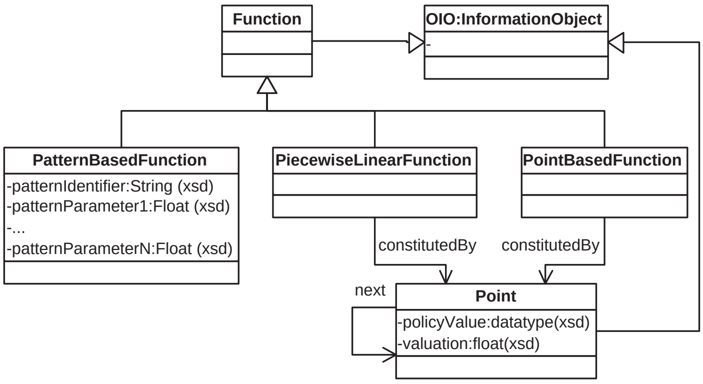


Figure 1: Modeling Valuation Functions

As depicted in Figure 1, a *Function*[7] is a specialization of *OIO:InformationObject* which represents abstract information existing in time and realized by some entity (Gangemi et al., 2004a). Our framework supports currently three ways of defining functions: *(i) Functions* can be modeled by specifying sets of points in $\mathbb{R}^2$ that explicitly map attribute values to valuations. This is particularly relevant for nominal attributes. *(ii)* We allow these points to be extended to piecewise linear value functions; this is important when dealing with continuous attribute values, as in the case of *Response Time*. *(iii)* Third, we allow the reuse of typical function patterns, which are mapped to predefined, parameterized valuation rules. Note that such patterns are not restricted to piecewise linear functions since all mathematical operators contained in the SWRL specification can be used. The different types of declarative modeling functions are discussed next in more detail.

(i) Point-based Functions

As illustrated in Figure 1, *PointBasedFunctions* are *Functions* that are *constitutedBy* a set of *Points*. Thus, the property *policyValue* refers to exactly one attribute value and the property *valuation* to exactly one utility measure that is assigned to this attribute value. Both properties are modeled by OWL datatypes. OWL datatypes mainly rely on the non-list XML Schema datatypes defined in Biron and Malhotra (2000). Depending on the attribute, *policyValue* either points to a *xsd:string*, *xsd:integer* or *xsd:float*. A *valuation* is represented by a *xsd:float* between 0 and 1 or by $-\infty$.

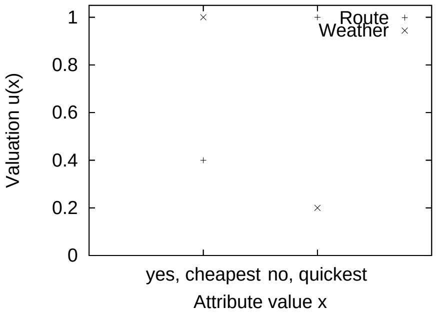


Figure 2: Example: Point-based value function

In our route planning example, a requester might specify her preferences with respect to the service property *Weather* by a *PointBasedFunction*, which is *constitutedBy* two instances of *Point* with $(''yes'', 1)$ and $(''no'', 0.2)$. This would indicate that the requester would highly prefer weather information to be taken into account, but has some minor use for routes calculated without it. Similarly, the preferences for the route calculation attribute can be defined with *Points* $(''quickest'', 1)$ and $(''cheapest'', 0.4)$. These mappings are illustrated in Figure 2.

(ii) Piecewise Linear Functions

In order to support the definition of *Functions* on continuous properties as well, we introduce *PiecewiseLinearFunctions* as shown in Figure 1. To express *PiecewiseLinearFunctions*, we extend the previous approach by the relation *next* between two *Points* with adjacent x-coordinates.

[7] Concepts and relations belonging to the ontology are written in *italics*. All concepts and relations imported from other ontologies are labelled with the corresponding namespace. Sometimes concept names in the text are used in plural to improve readability.

Such adjacent *Points* can be connected by straight lines, forming a piecewise linear value function as depicted in Figure 3. For every line between the *Points* (x_1, y_1) and (x_2, y_2) as well as a given *PolicyValue* x, we calculate the valuation v as follows:

$$v = \begin{cases} \frac{y_1-y_2}{x_1-x_2}(x-x_1)+y_1, & \text{if } x_1 \leq x < x_2 \\ 0, & \text{otherwise} \end{cases}$$

This equation is formalized using rule R1. To achieve this we exploit the math as well as the comparison built-in predicates provided by SWRL[8].

$$\begin{aligned} cal_v(v, x, x_1, y_1, x_2, y_2) \leftarrow\ & subtract(t_1, y_1, y_2), subtract(t_2, x_1, x_2), divide(t_3, t_1, t_2), \\ & subtract(t_4, x, x_1), multiply(t_5, t_3, t_4), add(v, t_5, y_1), \\ & lessOrEqualThan(x_1, x), lessThan(x, x_2) \end{aligned} \tag{R1}$$

As an example, let us assume the *Function* for the attribute *Response Time* of the route planing service is given by a *PiecewiseLinearFunction* with the *Points (0, 1)*, *(10, .8)*, *(30, .3)*, *(60, 0)* as shown in Figure 3. Now, we can easily find out to which *valuation v* a certain *policyValue x* is assigned. Thus, the predicate $cal_v(v, x, x_1, y_1, x_2, y_2)$ is true iff the *policyValue* x is between two adjacent *Points* (x_1, y_1) and (x_2, y_2) and the *valuation* equals v. For instance, for a *Response Time* of 20 seconds cal_v evaluates the straight line connecting the adjacent Points *(10, .8)* and *(30, .3)*, which results in a *valuation v* of .675.

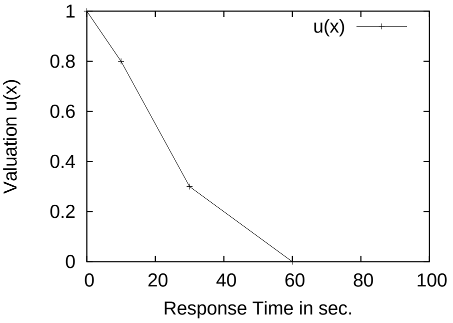


Figure 3: Example: Piecewise linear value function

(iii) Pattern-based Functions

Alternatively, value functions can be modeled by means of *PatternBasedFunctions.* This type refers to functions like $u_{p_1,p_2}(x) = p_1 e^{p_2 x}$, where p_1 and p_2 represent parameters that can be used to adapt the function. In our ontology, these *Functions* are specified through parameterized predicates which are identified by *patternIdentifiers.* A *patternIdentifier* is a *xsd:string* that uniquely refers to a specific SWRL-predicate. A *patternParameter* is a *xsd:float* that defines how a specific parameter of the pattern-predicate has to be set. To allow an arbitrary number of parameters in a rule, universal quantification over instances of *patternParameter* would be necessary. Since universal quantification in rule bodies is not expressible in SWRL, the different parameters are modeled as separate properties in the ontology, viz. *patternParameter1,. . . , patternParameterN.* Of course, this restricts the modeling approach, as the maximal number of

[8] For the sake of readability, we use a predicate with arity five. Techniques for reifying higher arity predicates are well known (Horrocks et al., 2000).

parameters has to be fixed at ontology design time. However, we believe that keeping the logic decidable justifies this limitation.

As shown in the example below (rule R2), each *pattern* is identified by a hard-coded internal string. Thus, in order to find out which *pattern*-predicate is applicable, the *patternIdentifer* specified in the policy is handed over to the *pattern*-predicate by using the first argument and then compared to the internal identifier. If the two strings are identical, the predicate is applied to calculate the corresponding *valuation* of a certain *policyValue*.

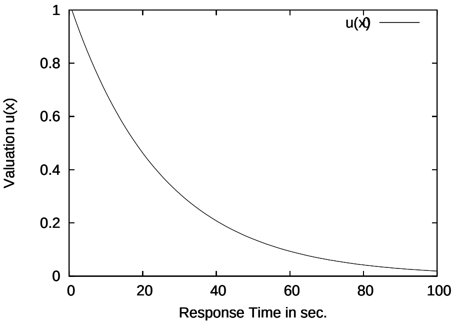


Figure 4: Example: Pattern-based Valuation Function

As an example, we again focus on the attribute *Response Time* of the route planning service. In many scenarios, the dependency between configurations and prices or valuations is given by functions. Assume the preferences for *Response Time* are given by the exponential function $u_{p_1,p_2}(x) = p_1 e^{p_2 x}$ with the *patternParameters* $p_1 = 1.03$ and $p_2 = -.04$ (Figure 4). Axiom R2 formalizes the pattern. The internal identifier in this example is *'id:exp'*. The corresponding comparison is done by the SWRL-built-in *equals*, which is satisfied iff the first argument is identical to the second argument. SWRL supports a wide range of mathematical built-in predicates (c.f. Horrocks et al., 2004) and thus nearly all functions can be supported.

$$\begin{aligned} \textit{pattern}(id, x, p_1, \ldots, p_n, v) \leftarrow\ & \textit{String}(id), \textit{PolicyValue}(x), \textit{Valuation}(v), \\ & \textit{equals}(id, \text{"}id:exp\text{"}), \textit{multiply}(t_1, p_2, x), \\ & \textit{pow}(t_2, \text{"}2.70481\text{"}, t_1), \textit{multiply}(v, p_1, t_2) \end{aligned} \tag{R2}$$

5.2 Modeling Policies and Configurations

As discussed in Section 3, we formalize user preferences as well as the provider's pricing information by means of policies. For instance, a price-conscious user might prefer a cheap service even if the service has a rather slow response time, whereas a time-conscious user might pay any price for a fast service. Hence, policies can be seen as different perspectives on a certain configuration. To model these policies, we use and specialize the DOLCE module Descriptions & Situations (DnS); doing so provides a basic theory of contextualization (Gangemi et al., 2003). Such a theory is required to reflect the fact that a certain configuration can be considered as more or less desirable depending on the scoring policies of a buyer or that a certain configuration can be priced at varying levels depending on the pricing policies of a seller.

When using DnS with DOLCE, we distinguish between DOLCE *ground entities*, which form a *DnS:Situation*, and *descriptive entities* composing a *DnS:Description*, i.e. the context in which *Situations* are interpreted. As depicted in Figure 5, we

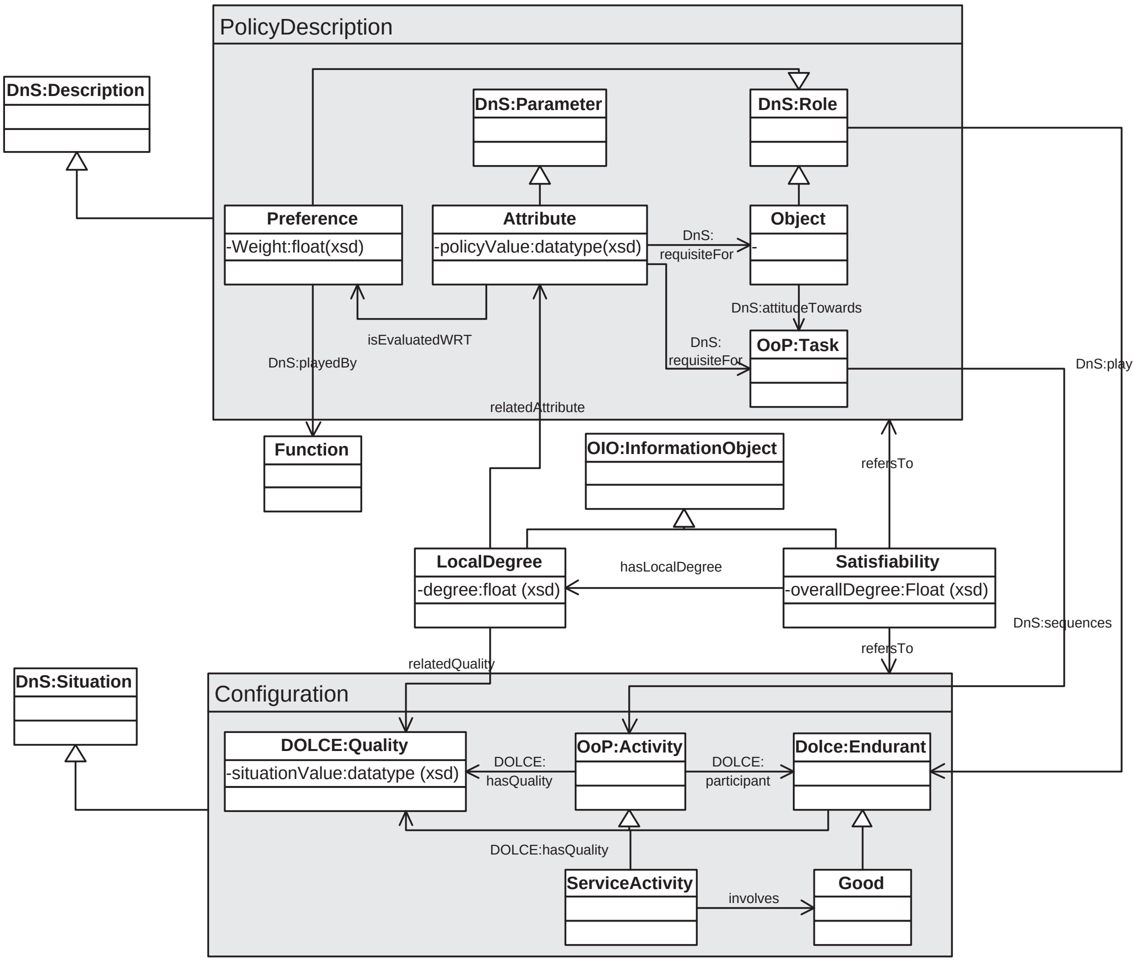


Figure 5: Policy description framework. To improve readability we illustrate certain relations by nesting UML classes: The class *PolicyDescription* has a *DnS:defines*-relation and the class *Configuration* possesses a *DnS:settingFor*-relation to each contained class.

specialize the *DnS:Description* to a *PolicyDescription* that can be used to evaluate concrete *Configurations* which are modeled as *Situations*. This distinction enables us, for example, to talk about products as roles on an abstract level, i.e. independent from the concrete entities playing the role. For instance, a certain product configuration can be evaluated in terms of either a seller's pricing policy or user preferences depending on the point of view.

In the following, we describe how such *Configurations* and *PolicyDescriptions* are modeled and then show how the evaluation of policies is carried out.

(i) Configuration

In a first step, we define the ground entities that describe a *DnS:Situation*. In our context, such *DnS:Situations* reflect configurations of concrete goods or services. Hence, we model *Configuration* as a subclass of *DnS:Situation* as shown in Figure 5. Since there are different ways of defining goods and services, a generic approach is used in this work; a concrete *Good* is represented by an instance of *DOLCE:Endurant* and a service by a combination of *DOLCE:Endurants* and *OoP:Activities* as done in Gangemi et al.

(2003). Specializations of *OoP:Activities* capture *ServiceActivities* like *RoutePlanning* while specializations of *DOLCE:Endurants* capture the corresponding objects *involved* in a *ServiceActivity* (such as inputs and outputs). Moreover, *DOLCE:Endurants* as well as *OoP:Activities* have *DOLCE:Qualities* with a datatype property of *situationValue*.

This means a concrete route planing service is represented by an instance of the following ontology: we specialize *ServiceActivity* to *PlanningActivity* with the *DOLCE:Qualities WeatherQuality*, *ResponseTimeQuality* and *AvailabilityQuality*. In addition, the *PlanningActivity involves* a *ServiceOutput* which specializes *Good*. *ServiceOutput* is associated to a *RouteQuality* that defines whether the output is the cheapest or the quickest route. Note that a *DOLCE:Quality* in a concrete configuration has exactly one *situationValue*-property.

(ii) Policy Description

In a second step, the descriptive entities are specialized in order to define policies that can be used to specify buyer scoring functions as well as pricing functions of configurable trading objects. As shown in Figure 5, policies are modeled as a specialization of *DnS:Description* called *PolicyDescription*. Here the relation *DnS:defines* associates a *DnS:Role* representing the *Object* on which the policy is defined; this could be a certain type of good or the output of a service. Since they are modeled as *DnS:Roles*, policies can be defined on an abstract level without referring to a concrete *DOLCE:Endurant*. For instance, preferences can be defined for a certain product category (such as computers) as a whole and not only for one concrete computer system. Furthermore, a *PolicyDescription* could also regulate an *OoP:Task*. This is for example the case when talking about Web services. A route planing service might execute a *RoutePlanningTask* where the *Input* (specialization of *Object*) is a certain destination and the *Output* (specialization of *Object*) is the calculated route.

However, as discussed in Section 3, the configurations are preferred to varying degrees depending on the concrete properties of the trading object. We model this by introducing the *DnS:Parameter Attribute*, which is a *DnS:requisiteFor* an *Object* or *OoP:Task*. *Attributes* have a datatype property *policyValue* pointing to all possible attribute values. Furthermore, each *Attribute* is assigned to a certain preference structure. As discussed above, preference structures on attributes are imposed by *Functions*. *Functions* are *OIO:InformationObjects* (cf. Figure 1) that play the role of *Preferences* in a *PolicyDescription* and define how *policyValues* are mapped to *valuations*. In other words, a policy defines which *Function* should be used in which context (i.e. for which attribute). Besides *Functions*, *Preferences* also define the relative importance of the given *Attribute*.

After discussing how *Configurations* as well as *PolicyDescriptions* are modeled, we introduce the rules for evaluating concrete *Configurations* with respect to given policies. We thus show how pricing policies and scoring policies are applied to determine the price of a configuration and willingness to pay, respectively.

(iii) Policy Evaluation

With our approach, policies that contain *Functions* no longer lead only to a pure boolean statement about the conformity of a *Configuration*, but rather to a degree of conformity of the *Configuration.* Therefore, the traditional *DnS:satisfies*-relation between a *DnS:Situation* and *DnS:Description* stemming from DOLCE is no longer sufficient since additional information about the degree of conformity has to be captured. Ontologically, this requires putting in relation the *PolicyDescription*, a concrete *Configuration* and an *overallDegree* that represents the valuation in which the latter satisfies the former. As tertiary relations cannot be modeled with the formalism at hand directly, the *OIO:InformationObject Satisfiability* is introduced to link the three entities. We use the relation *refersTo* to identify the *PolicyDescription* as well as the *Configuration* for which the datatype property *overallDegree* represents the valuation. Figure 5 sketches this modeling approach.

In line with the utility model defined in Equation 3, we first calculate the valuation for each attribute individually and then aggregate the individual valuations to the *overallDegree.* To represent individual valuations for attributes we introduce the concept *LocalDegree*, which is also an *OIO:InformationObject* in terms of DOLCE. *LocalDegree* links each *DOLCE:Quality* of the *Configuration* to a certain *Attribute* in the *PolicyDescription* by using the associations *relatedAttribute* as well as *relatedQuality.*

On this basis, the property *degree*, which can be interpreted as the valuation a single attribute contributes to the overall valuation, is calculated as follows: depending on which *Function DnS:plays* the role of *Preference* for a certain *Attribute*, one of the rules below is used. In order to abbreviate the following rule definitions, we first define the shortcut relation *isDeterminedBy* between a *LocalDegree* and the *Function* that specifies the *Preferences* for the *Attribute* related to the *LocalDegree.*

$$\begin{aligned} \mathit{isDeterminedBy}(x, f) \leftarrow {} & \mathit{relatedAttribute}(x, a), \mathit{isEvaluatedWRT}(a, p), \\ & \mathit{DnS{:}playedBy}(p, f) \end{aligned} \tag{R3}$$

In case of *PointBasedFunctions* we look up the *situationValue* in the *Configuration* and compare this value with the *policyValues* of all *Points* defined by the *Function.* If the *policyValue* of a *Point* p matches the *situationValue*, the property *valuation* of p determines the *degree.* For instance, if the *situationValue* for *RouteQuality* is determined by the *xsd:string* "quickest" and the *PointBasedFunction* is given by $(\text{"}\mathit{quickest}\text{"}, 0.6)$ and $(\text{"}\mathit{cheapest}\text{"}, 0.4)$, we derive a *degree* of 0.6 since only the *policyValue* "quickest" matches the *situationValue.* This is formalized in Rule R4. Correspondingly, the Rules R5 and R6 can be used to evaluate *PiecewiseLinearFunctions* and *PatternBasedFunctions*, respectively. Rule R5 uses the cal_v-predicate (defined in Rule R1) and Rule R6 employs the *pattern*-predicate (defined in Rule R2) to determine allowed mappings between *policyValues* and *valuations.*

$$
\begin{aligned}
\mathit{degree}(ld, v) \leftarrow\ & \mathit{isDeterminedBy}(ld, f), \boldsymbol{\mathit{PointBasedFunction}}(f), \mathit{constitutedBy}(f, po),\\
& \mathit{policyValue}(po, pv), \mathit{relatedQuality}(ld, q), \mathit{situationValue}(q, y),\\
& \mathit{equals}(pv, y), \mathit{valuation}(po, v) \qquad \text{(R4)}\\
\mathit{degree}(ld, v) \leftarrow\ & \mathit{isDeterminedBy}(ld, f), \boldsymbol{\mathit{PiecewiseLinearFunction}}(f),\\
& \bigwedge_{i\in\{1,2\}} (\mathit{constitutedBy}(f, po_i), \mathit{policyValue}(po_i, pv_i), \mathit{valuation}(po_i, v_i)),\\
& \mathit{next}(po_1, po_2), \mathit{relatedQuality}(ld, q), \mathit{situationValue}(q, y),\\
& \mathit{cal}_v(v, y, pv_1, v_1, pv_2, v_2) \qquad \text{(R5)}\\
\mathit{degree}(ld, v) \leftarrow\ & \mathit{isDeterminedBy}(ld, f), \boldsymbol{\mathit{PatternBasedFunction}}(f), \mathit{patternIdentifier}(f, id),\\
& \bigwedge_{i\in\{1,\ldots,n\}} (\mathit{patternParameter}_i(f, p_i)), \mathit{relatedQuality}(ld, q),\\
& \mathit{situationValue}(q, y), \mathit{pattern}(id, y, p_1, \ldots, p_n, v) \qquad \text{(R6)}
\end{aligned}
$$

The valuation derived from the individual attributes is weighted according to their relative importance defined by the concept *Weight*. The weights λ_i of the individual attributes i have to be normalized between 0 and 1. This is done by means of the formula $\frac{1}{n}\sum_{i=1}^{n}\lambda_i$, which is evaluated within rule R7. Based on the normalized weights, Rule R7 calculates the *overallDegree* by encoding equation 1.

$$
\begin{aligned}
\mathit{overallDegree}(s, v) \leftarrow\ & \mathit{PolicyDescription}(d), \mathit{refersTo}(s, d), \mathit{Configuration}(c), \mathit{refersTo}(s, c),\\
& \bigwedge_{i\in\{1,\ldots,n\}} (\mathit{hasLocalDegree}(s, ld_i), \mathit{isDeterminedBy}(ld_i, f_i),\\
& \mathit{weight}(f_i, w_i)), \mathit{sum}(g, w_1, \ldots, w_n), \bigwedge_{i\in\{1,\ldots,n\}} (\mathit{div}(r_i, w_i, g),\\
& \mathit{localDegree}(ld_i, v_i), \mathit{mul}(k_i, v_i, r_i)), \mathit{sum}(v, k_1, \ldots, k_n) \qquad \text{(R7)}
\end{aligned}
$$

As an example, we assume the *Configuration s* of a route planning service, which returns the quickest route while considering traffic information. Further, a response time of 20 seconds is guaranteed. Based on the example *Functions* above, this leads to local *degrees* of 1 for the *Attribute Weather*, 1 for *Traffic*, 0.47 for *Response Time* and 0.5 for *Route*. Moreover, we assume that the PolicyDescription contains the weights of 2, 2, 1, 1 for the individual *Attributes*. Now we can query the knowledge base containing the configuration c as well as the PolicyDescription d (e.g. by using SPARQL[9]) to obtain the *overallDegree* of a *Satisfiability* instance that *refersTo c* and d. In the example, this would result in a *overallDegree* of 0.83. In addition, configurations can be automatically classified by concepts according to their *overallDegree* (e.g. all configurations with a *overallDegree* above 0.8 according to a certain Scoring Policy). Such automatic classification of configurations is particularly important for implementing context-aware decision-making. For example, different contexts (current location, time, etc.) may require different scoring policies. In our rule-based framework, such *context-rules* can be easily added to the knowledge base.

[9] http://www.w3.org/TR/rdf-sparql-query/

6 Conclusion

In this paper, we provided a formal representation of a utility-based policy framework which realizes the advantages of *Utility Function policies*, such as preference modeling and inherent conflict resolution, with a purely declarative and standard-based approach. This is essential for flexibility and interoperability of the system within electronic markets. As a use case, it was shown how such a policy framework can be applied for the formal modeling of a provider's pricing policies as well as requester preferences. We believe that expressing the relations between product/service configurations and prices or the willingness to pay is crucial in electronic markets for configurable trading objects. To the best of our knowledge, there are no formal, declarative and standard-based languages available yet that provide sufficient expressivity to support this.

In a next step, we plan to integrate the policy framework into a larger ontology for expressing offers, requests and agreements in a market. In this context, super- as well as subadditive valuations should be supported through primitives for representing *bundles* and *substitutes*. Such a bidding language could be used to express requests and offers in a multi-attribute electronic market. Hence, the ontology has to be mapped to an allocation problem, which efficiently encodes the winner and price determination in the market based on pricing as well as scoring policies.

References

Anderson A. et al. (2003): "XACML Profile for Web Services," `http://xml.coverpages.org/WSPL-draft-xacmlV04-1.pdf`.

Ashley, P. et al. (2003): "Enterprise Privacy Authorization Language," W3C Submission, `http://www.w3.org/Submission/EPAL`.

Baader, F., D. Calvanese, D. McGuinness, D. Nardi, and P. F. Patel-Schneider (eds.) (2003): *The Description Logic Handbook: Theory Implementation and Applications*, Cambridge University Press.

Bichler, M. and J. Kalagnanam (2005): "Configurable offers and winner determination in multi-attribute auctions," *European Journal of Operational Research*, 160(2), pp. 380–394.

Biron, P. V. and A. Malhotra (2000): "XML Schema Part 2: Datatypes," W3C Recommendation. Latest version is available at `http://www.w3.org/TR/xmlschema-2/`.

Boutilier, C. and H. H. Hoos (2001): "Bidding Languages for Combinatorial Auctions," in: *International Joint Conference on Artificial Intelligence IJCAI'01*, pp. 1211–1217.

Brockmans, S., R. Volz, A. Eberhart, and P. Löffler (2004): "Visual modeling of OWL DL ontologies using UML," in: S. M. et al. (ed.), *Proc. of the 3rd International Semantic Web Conference*, Springer LNCS, Hiroshima, Japan, pp. 198–213.

Gangemi, A., S. Borgo, C. Catenacci, and J. Lehmann (2004a): "Task taxonomies for knowledge content," Metokis deliverable d07.

Gangemi, A., P. Mika, M. Sabou, and D. Oberle (2003): "An Ontology of Services and Service Descriptions," working paper, Laboratory for Applied Ontology, Rome, Italy.

Gangemi, A., M.-T. Sagri, and D. Tiscornia (2004b): "A Constructive Framework for Legal Ontologies," Internal project report, EU 6FP METOKIS Project, Deliverable, `http://metokis.salzburgresearch.at`.

Grid Resource Allocation and Agreement Protocol Working Group (2005): "Web Services Agreement Specification," `https://forge.gridforum.org/projects/graap-wg/document/WS-AgreementSpecification/en/7`.

Grosof, B. and T. Poon (2003): "SweetDeal: Representing agent contracts with exceptions using XML rules, ontologies, and process descriptions," in: *Proceedings of the 12th International Conference on the World Wide Web (WWW 2003)*, Budapest, Hungary.

Gruber, T. R. (1993): "A translation approach to portable ontologies," *Knowledge Acquisition*, 5(2), pp. 199–220.

Horrocks, I. and P. F. Patel-Schneider (2004): "A proposal for an OWL rules language," in: *Proceedings of the 13th International Conference on the World Wide Web (WWW 2004)*, ACM Press, New York, USA, pp. 723–731, doi:http://doi.acm.org/10.1145/988672.988771.

Horrocks, I., P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean (2004): "SWRL: A Semantic Web Rule Language Combining OWL and RuleML," W3C Submission, available at `http://www.w3.org/Submission/SWRL`.

Horrocks, I., P. F. Patel-Schneider, and F. van Harmelen (2003): "From SHIQ and RDF to OWL: The making of a web ontology language," *Journal of Web Semantics*, 1(1), pp. 7–26.

Horrocks, I., U. Sattler, S. Tessaris, and S. Tobies (2000): "How to decide query containment under constraints using a description logic," in: *Proceedings of the 7th International Conference on Logic for Programming and Automated Reasoning (LPAR'2000)*.

Kagal, L. (2004): *A Policy-Based Approach to Governing Autonomous Behavior in Distributed Environments*, Ph.D. thesis, University of Maryland Baltimore County, Baltimore MD 21250.

Keeney, R. L. and H. Raiffa (1976): *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, J. Wiley, New York.

Kephart, J. O. and W. E. Walsh (2004): "An Artificial Intelligence Perspective on Autonomic Computing Policies," in: *Proc. of 5th IEEE Int. Workshop on Policies for Distributed Systems and Networks*, IEEE Computer Society, Yorktown Heights, New York, USA, pp. 3–12.

Lamparter, S., A. Eberhart, and D. Oberle (2005): "Approximating Service Utility from Policies and Value Function Patterns," in: *6th IEEE Int. Workshop on Policies for Distributed Systems and Networks*, IEEE Computer Society, Stockholm, Sweden, pp. 159–168.

Malone, T. W., K. Crowston, B. P. J. Lee, C. Dellarocas, G. Wyner, J. Quimby, C. S. Osborn, A. Bernstein, G. Herman, M. Klein, and E. O'Donnell (1997): "Tools for Inventing Organizations: Toward a Handbook of Organizational Processes," *Management Science*, 45(3).

Masolo, C., S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, and L. Schneider (2002): "The WonderWeb Library of Foundational Ontologies," WonderWeb Deliverable D17, `http://wonderweb.semanticweb.org`.

Motik, B., U. Sattler, and R. Studer (2005): "Query Answering for OWL-DL with Rules," *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 3(1), pp. 41–60.

Neumann, D. (2004): *Market Engineering - A Structured Design Process for Electronic Markets*, Ph.D. thesis, Department of Economics and Business Engineering, University of Karlsruhe (TH), Karlsruhe.

Nisan, N. (2000): "Bidding and allocation in combinatorial auctions," in: *Proceedings of the 2nd ACM conference on Electronic commerce (EC'00)*, ACM Press, New York, NY, USA, pp. 1–12.

Russel, S. and P. Norvig (2003): *Artificial Intelligence - A Modern Approach*, second edition, Prentice Hall Series in Artificial Intelligence.

Sakellariou, R. and V. Yarmolenko (2005): "On the flexibility of WS-agreement for job submission," in: *Proceedings of the 3rd Intern. Workshop on Middleware for Grid Computing (MGC'05)*, ACM Press, New York, NY, USA, pp. 1–6.

Siddharth Bajaj et al. (2004): "Web Services Policy Framework," `http://www-128.ibm.com/developerworks/library/specification/ws-polfram`.

Uszok, A., J. M. Bradshaw, and R. Jeffers (2004): "KAoS: A Policy and Domain Services Framework for Grid Computing and Semantic Web Services," in: *Trust Management: Second International Conference, iTrust 2004, Oxford, UK, March 29 - April 1, 2004. Proceedings*, *LNCS*, vol. 2995, Springer, pp. 16–26.

W3C (2004): "Web Ontology Language (OWL)," `http://www.w3.org/2004/OWL/`, w3C Recommendation.

Pooling Uncertainty in a Permit Trading System: An Incentive for Collusion?

Ralf Löschel[1], Siegfried Berninghaus[1], and Jürgen Kühling[2]

[1] Institute for Economic Theory and Operations Research (WIOR),
Universität Karlsruhe (TH)
{loeschel,berninghaus}@wiwi.uni-karlsruhe.de

[2] Institute of Information Law,
Universität Karlsruhe (TH)
kuehling@ira.uka.de

Summary. This paper investigates the impact of stochastic emissions on a permit trading system and the effect of "pooling" firms' permits market activities. To this end, we develop a game theoretic model to show that in a duopoly industry, firms overinvest in emissions abatement and reduce output when emissions are uncertain. To mitigate these undesirable effects, the European emissions trading law allows firms to pool on the permits market. We demonstrate that pooling leads to less emissions abatement and higher output, but also enables firms to collude on the product market via the permits market. Since emissions allowances are an essential production input, firms can restrict their output by agreeing to emit less. A legal analysis shows that abusing the pooling-option for collusion conflicts EC competition law.

1 Introduction

Since early 2005, most European companies with carbon-dioxide-intensive production have been subject to the European emissions trading system, which obliges them to cover their emissions with permits. These permits, sometimes called allowances, are transferable between firms and other traders. Companies are thereby allowed to "pool" their activities on the permits market under the European emissions trading law (ETL) due to e.g. uncertainty-decreasing effects.

By means of a game theoretic analysis we first investigate the impact of uncertainty on emissions trading for an oligopolistic product market. In the main part of the paper we show that on the one hand, the "pooling option" diminishes undesirable uncertainty effects, but on the other hand creates a loophole for collusion on the output market via the permits market. Thus, pooling between oligopolists might be prohibited in order to avoid deadweight loss.

Article 28 of the European Directive on Emission Permits Trading (2003/87/EC) allows member states to give installations from the same activity the opportunity to pool. In other words, only firms from the same industry are able to jointly trade permits at the market.[1] The operators of the installations are required to name a trustee responsible for trading permits and covering emissions in the name of all participating operators. The trustee is also in charge in case emissions of the pool are uncovered.

The "pooling option" was introduced by German industry lobbying efforts. The intention was to preserve its voluntary commitment to reduce CO_2 emissions by establishing mandatory pools for all firms within a given industry. In this scenario, firms jointly reduce their total emissions and act on the permits market as a single trader. However, the German industry only succeeded in establishing a voluntary pooling option. Hence, as pools are not obligatory for all firms and permits sellers are disadvantaged in such pools, it is expected that few industries will make use of this option.[2] Nevertheless, most EU-Member States have enacted pooling into national law and some declarations of intent have already been made.[3] By Article 24 TEHG Germany also allows pooling for firms participating in emissions trading which is reasoned with the possibility to reduce administrative costs and the negative effects of uncertainty.[4]

Concerning the existing literature on uncertainty in permits trading systems, most contributions analyze uncertain amounts of ex-post emissions, as future emissions from any given installation are difficult to predict with certainty. Fluctuations in electricity demand, purity of raw materials or mechanical breakdowns can have an impact on the creation of stochastic forecasts. Even if emissions are not uncertain, measurement errors in monitoring can cause stochastic penalties. Carlson and Sholtz (1994) compare different types of issue and expiration dates for permits. They reason that a permit trading system like the European one is not efficient if emissions are uncertain; for example, in the case of high non-compliance penalties, excessive permit holdings (compared to expected emissions) emerge as "insurance" against falling short.

Mrozek and Keeler (2004) compare non-tradeable and tradeable permits with uncertain emissions levels. They find out that emissions are always closer to the optimum when permits are tradeable. Hennessy and Roosen (1999) analyze the effects of merging of firms when emissions are uncertain. They argue that under perfect competition, expected profits for merged firms are at least as high as the sum of expected profits for single firms. Moreover emissions are closer to the cost efficient solution when firms purchase permits together. Our paper is mainly based on the work of Mrozek and Keeler (2004), Hennessy and Roosen (1999) as well as on the model of Ehrhart et al. (2006).

In contrast to the models described above, Maeda and Tezuka (2004) analyze uncertain future emissions in a system with intertemporal trade. They show a negative impact of uncertainty on the present permit price. Ben-David et al. (2000) explore uncertain permit prices in an experimental laboratory setting. Montero (1997) assumes

[1] See Spieth and Röder-Persson (2003).
[2] See Klinski (2003) and Graichen and Requate (2005).
[3] See Betz et al. (2004).
[4] See Betz et al. (2003).

uncertainty in a way that allows for some trades to actually be interdicted by the regulator in order to e.g. avoid emission hot-spots. This results in an overall reduced welfare. Newell and Pizer (2000) compare tax and tradeable permits systems when abatement cost functions are stochastic.

Our paper is structured as follows: In Section 2.1, the effect of uncertain emissions is analyzed for an oligopolistic industry. Section 2.2 shows that pooling activities on the permits market reduce negative uncertainty effects both on output and the emissions level. On the other hand, firms have no incentive to found such pools, as their expected profits decrease. In Section 2.3 we assume, similar to Ehrhart et al. (2006), that firms abuse pooling in order to agree on emissions, i.e. firms eventually restrict output to increase returns. Thereby the market outcome can be worse than without pooling. Section 2 concludes with some possible extensions of the model. In Section 3 we briefly analyze the results of this paper from a legal perspective concerning European ETL and discuss whether pooling conflicts antitrust law. The paper ends with a brief conclusion.

2 Model

We analyze uncertain emission levels in a duopolistic industry with a simple theoretic model. Two risk-neutral firms, 1 and 2, produce a homogeneous commodity. Each firm's output is denoted as $y_i > 0$ ($i \in \{1,2\}$). The firms are Cournot competitors facing a linear monotonically decreasing inverse demand function $P(Y)$,[5]

$$P(Y) = P(y_i + y_{-i}), P'(Y) < 0.$$

As a fallout of production, every firm i generates expected emissions $e_i > 0$ that should be covered by holding permits $q_i \geq 0$ on hand. In the beginning, every firm i gets a free initial endowment of permits $q_i^0 \geq 0$. Firms can sell superfluous permits or buy more if they run short. Permits are traded on an allowances market on which firms are price takers, and hence take the permit price $p > 0$ as exogenously given. The costs for buying or gains from selling permits are $p \cdot (q_i - q_i^0)$.

In addition to permits expenses (or gains), firms also face production costs. We assume symmetric firms, i.e. both firms have the same cost function $C_i(\cdot) = C(y_i, e_i)$ $\forall i \in \{1,2\}$, depending on output y_i and emissions e_i. Producing the same output y_i with less emissions means investing in abatement measures, i.e. production costs increase. The cheapest abatement measures are always realized first. Therefore, the cost function $C(\cdot)$ is convex and decreases in emissions,

$$\frac{\partial C(y_i, e_i)}{\partial e_i} < 0 \;\wedge\; \frac{\partial^2 C(y_i, e_i)}{\partial e_i^2} > 0.$$

Since the non-abated emissions level is not determined ex-ante, uncertainty arises from the ex-post amount of emissions, $\bar{e_i} = e_i + e_i\theta_i$, whereas θ_i is a random variable with

[5] The index $-i$ represents the competitor of player i.

support on the interval $[-a, a]$, $0 < a < 1$. θ_i is symmetrically distributed with equal density $g(\theta_i)$ for both firms, $E(\theta_i) = 0$. $G(\theta_i)$ denotes the cumulative distribution function. θ_1 and θ_2 are assumed to be independent.

2.1 Effects of Uncertain Emissions

First we analyze the effect of uncertainty when firms simultaneously choose their amount of output, emissions and permits. In case firm's ex-post emissions exceed its permits holdings ($e_i + e_i\theta_i - q_i > 0$) it has to pay a fine f for every uncovered emission unit. We assume $f > p$; therefore, firms paying the fine are never better off than firms buying permits. The expected fine for firm i is $f \cdot E\{max[e_i + e_i\theta_i - q_i; 0]\}$.

A firm's expected profit Π_i^S is:

$$\begin{aligned} &E\{\Pi_i^S(y_i, e_i, q_i)\} = \\ &= P(Y)y_i - C(y_i, e_i) - p(q_i - q_i^0) - f \cdot E\{max[e_i + e_i\theta_i - q_i; 0]\} \\ &= P(Y)y_i - C(y_i, e_i) - p(q_i - q_i^0) - f \cdot \textstyle\int_{(q_i-e_i)/e_i}^{a}(e_i + e_i\theta_i - q_i)dG(\theta_i) \end{aligned}$$

S symbolizes that each firm maximizes its own expected profits by setting y_i, e_i and q_i optimally. The corresponding first order conditions are

$$P(Y) + P'(Y)y_i - C_{y_i}(y_i, e_i) = 0, \tag{1}$$

$$-C_{e_i}(y_i, e_i) - f \cdot [1 - G((q_i - e_i)/e_i) + \int_{(q_i-e_i)/e_i}^{a} \theta_i dG(\theta_i)] = 0, \text{ and} \tag{2}$$

$$-p + f \cdot [1 - G((q_i - e_i)/e_i)] = 0 \tag{3}$$

for all $i \in \{1, 2\}$.[6]

In case the corresponding Hesse-matrix for every player i is negative semidefinite, a tuple (y_i^*, e_i^*, q_i^*) satisfying the first order conditions maximizes a firm's expected profit (given the other firm's actions).

Since $g(\theta_i)$ is symmetric around the expected value 0, $G(0) = 1/2$ and by equation (3)

$$\begin{aligned} q_i &< e_i \text{ if } f < 2p, \\ q_i &= e_i \text{ if } f = 2p, \text{ and} \\ q_i &> e_i \text{ if } f > 2p. \end{aligned}$$

If $f = 2 \cdot p$, the firms' expected emissions equal their purchased amount of allowances. In case of a relatively high fine $f > 2p$, firms buy too many permits in order to avoid a penalty. If the fine f is less than $2p$, firms speculate for fewer emissions than expected and therefore buy fewer permits.[7]

[6] We define $F_X(X, Y)$ as the (partial) derivative of the function $F(X, Y)$.

[7] This result coincides with the result by Mrozek and Keeler (2004).

Substituting (3) into (2) yields

$$-C_{e_i}(y_i, e_i) - p - p \cdot E[\theta_i|\theta_i \geq (q_i - e_i)/e_i] = 0.$$

In case $g(\theta_i)$ is positive on $[\frac{q_i - e_i}{e_i}, a]$, marginal abatement costs $-C_{e_i}(y_i, e_i)$ are higher than the permit price p. It is a well-known result in the literature that under perfect competition a necessary condition for a cost-minimizing emissions trading system is the equality of permit price and marginal abatement costs.[8] Taking this as a reference point, uncertain ex-post emissions do not lead to the cost efficient solution, as the firms' marginal abatement costs are too high. Further we find out that—despite uncertain emissions— the equilibrium solution is independent of the initial endowment q_i^0, which is well-known for basic models with certain ex-post emissions.

For the sake of simplification and comparability of our results, we have to make some more precise assumptions. First, we assume that a regulator wants the firms to emit as many emissions as he has issued permits ($q_i = e_i$).[9] Therefore, as stated above, $f = 2p$. Second, we consider θ_i as uniformly distributed on $[-a, a]$, with $0 < a < 1$. Additionally, we define the cost function as $C(y_i, e_i) = c[(y_i - e_i)^+]^2$ and the inverse demand function as $P(y_i + y_{-i}) = m - t(y_i + y_{-i})$. Since equilibrium is not dependent on initial endowment, we assume that firms receive no free permits in the beginning and have to buy all permits on the permits market, $q_i^0 = 0$.

Hence, for the first order conditions (1), (2) and (3) we obtain

$$m - t(2y_i + y_{-i}) - 2c(y_i - e_i) = 0,$$

$$2c(y_i - e_i) - 2p(\frac{(1+a)^2}{4a} - \frac{q_i^2}{4ae_i^2}) = 0, \text{ and}$$

$$-p(\frac{q_i - e_i}{ae_i}) = 0$$

for all $i \in \{1, 2\}$, given the other firm's actions. The second order conditions for a maximum are satisfied, as the corresponding Hessian is negative semidefinite for all $i \in \{1, 2\}$. Thus, the equilibrium solution (y_i^*, e_i^*, q_i^*) is

$$y_i^* = \frac{2m - (2+a)p}{6t}, \text{ and}$$

$$e_i^* = q_i^* = \frac{2m - (2+a)p}{6t} - \frac{(2+a)p}{4c}$$

for all $i \in \{1, 2\}$.

[8] See e.g. Montgomery (1972).

[9] If firms buy fewer permits than their expected emissions, the expected total emissions of the whole trading system will exceed the amount of permits. In this case, the reduction goal set by the regulator is not achieved. The other way round, i.e. if the fine f exceeds a certain threshold, then expected total emissions in the system are lower than the amount of permits. In that case the reduction goal is overachieved. We assume that the regulator wants as many emissions as he issued permits.

To do some comparative statics with regard to uncertainty, we increase the mean preserving spread of θ_i. In our simplified framework this corresponds to an increase of a.

$$\frac{dy_i^*}{da} = -\frac{p}{6t} < 0$$

$$\frac{de_i^*}{da} = \frac{dq_i^*}{da} = -\frac{p}{6t} - \frac{p}{4c} < 0$$

The derivations show that increasing a leads to lower emissions e_i^* and therefore to lower outputs y_i^*. Because of the assumption $f = 2p$, the effect of changing a on permits q_i^* equals the effect on emissions e_i^*.[10]

The expected profit of firm i is

$$E\{\Pi_i^S\}^* = \frac{4c(2m-(2+a)p)^2 + 9(2+a)^2p^2t}{144ct}.$$

Recapitulating, increasing uncertainty leads to decreasing output, decreasing emissions and decreasing permit holdings. Lower outputs cause a higher product price, affecting consumers' surplus in a negative way. Lower emissions represent overly high investments in emissions abatement compared to the abatement costs-minimizing solution. Thus, the regulator should be interested in reducing uncertainty in an emissions trading system.

2.2 Pooling

In order to achieve cost-efficient emissions abatement and to avoid negative effects on the product market, the regulator should reduce the emissions uncertainty that firms face. One possibility might be a "pooling option", which is included in the European emissions trading law. Using this option, firms transfer their permit holdings to a trustee, who is authorized to act on the allowance market. The trustee is also responsible for uncovered emissions.

Should both firms in Section 2.1 use the pooling option, they only have to pay a fine if their total amount of permits Q is less than their total emissions $\overline{e_1} + \overline{e_2}$. In order to calculate the expected emissions and the expected fine after pooling, we define for each firm the random variable $X_i := e_i\theta_i$. It denotes the ex-post deviation from expected emissions e_i and is uniformly distributed on the interval $[-e_ia; e_ia]$. The firms' distribution of the cumulated deviation from the cumulated expected emissions is calculated by the convolution $Z = X_1 + X_2$. Without loss of generality, assume $e_i \geq e_{-i}$. Then the density of Z is:

[10] The results for y_i^* and e_i^* hold independent of f. Only the effect of increasing a on q_i^* is ambiguous for different values of f.

$$h(Z) = \begin{cases} \frac{Z+a(e_i+e_{-i})}{4a^2 e_i e_{-i}} & \text{if } Z \in [-(e_i+e_{-i})a; (e_{-i}-e_i)a) \\ \frac{1}{2ae_i} & \text{if } Z \in [(e_{-i}-e_i)a; (e_i-e_{-i})a] \\ \frac{-Z+a(e_i+e_{-i})}{4a^2 e_i e_{-i}} & \text{if } Z \in ((e_i-e_{-i})a; (e_i+e_{-i})a] \\ 0 & \text{else} \end{cases}$$

Expected total pool costs F^P are[11]

$$\begin{aligned} E\{F^P\} = & \quad p \cdot Q + f \cdot E\{max[e_1+e_2+Z-Q; 0]\} \\ & = p \cdot Q + f \cdot \int_{Q-e_1-e_2}^{a(e_1+e_2)} (e_1+e_2+Z-Q)h(Z)dZ. \end{aligned}$$

We assume that the amount of total permits Q is chosen (by the trustee) in a way that minimizes expected pool costs. Choosing Q optimally implies

$$p - f \cdot [1 - H(Q-e_1-e_2)] = 0 \tag{4}$$

The corresponding second order condition for the choice of Q is

$$f \cdot h(Q-e_1-e_2) \geq 0.$$

Since we assume $f = 2p$, (4) implies $H(Q-e_1-e_2) = 1/2$, where $H(\cdot)$ denotes the corresponding c.d.f. Hence, as $h(\cdot)$ is symmetric and $E(Z) = 0$, the pool covers the cumulated expected emissions ($Q = e_1 + e_2$). Furthermore, we assume that both firms anticipate the cost-minimizing behavior of the trustee.

Obviously, the distribution of the pool costs F^P among the firms is a crucial question for pool formation. Firms must agree on a cost allocation rule, where ϕ_i denotes firm i's share. Trivially, the sum of all shares must equal one ($\phi_i + \phi_{-i} = 1$). As the firms are symmetric in our simplified setting, assume that firms divide costs equally, $\phi_i = 1/2$.[12] Including the sharing rule and the behavior of the trustee ($Q = e_1 + e_2$), firm i's expected profit is

$$E\{\Pi_i^P\} = P(Y)y_i - C(y_i, e_i) - [p(e_1+e_2) + f \int_0^{a(e_1+e_2)} Zh(Z)dZ]/2.$$

As above, we assume that every firm i chooses its output y_i and its expected emissions level e_i in a profit-maximizing way.

Choosing y_i and e_i optimally implies

$$0 = m - t(2y_i + y_{-i}) - 2c(y_i - e_i), \tag{5}$$

[11] The exponent P indicates that firms use the pooling option.

[12] In reality (where symmetry can hardly be found) the cost allocation question is essential. It is supposedly the result of intense bargaining. From a game theoretic point of view, the well-known cooperative solution concepts provide a toolbox for establishing a "reasonable" allocation pattern.

$$0 = \begin{cases} 2c(y_i - e_i) - \frac{p}{2} - \frac{pa(3{e_i}^2 - {e_{-i}}^2)}{12{e_i}^2} & \text{if } e_i > e_{-i} \\ 2c(y_i - e_i) - \frac{p}{2} - \frac{pae_i}{6e_{-i}} & \text{if } e_i \leq e_{-i} \end{cases} \tag{6}$$

for all $i \in \{1, 2\}$.

The symmetric solution y_i^{**}, e_i^{**} and Q^{**}, satisfying (5), (6) and (4), is

$$y_i^{**} = \frac{6m - (3+a)p}{18t},$$

$$e_i^{**} = \frac{6m - (3+a)p}{18t} - \frac{(3+a)p}{12c}, \text{ and}$$

$$Q^{**} = \frac{6m - (3+a)p}{9t} - \frac{(3+a)p}{6c}$$

for all $i \in \{1, 2\}$.

The sufficient conditions for y_i^{**} and e_i^{**} to establish a maximum are fulfilled because the corresponding Hessian is negative semidefinite for all $i \in \{1, 2\}$.

Firm i's expected profit is

$$E\{\Pi_i^P\}^{**} = \frac{8c(18m^2 - 15(3+a)mp + 2(3+a)^2p^2) + 27(3+a)^2p^2t}{1296ct}$$

for all $i \in \{1, 2\}$.

As one can easily see, $e_i^{**} > e_i^*$ and $y_i^{**} > y_i^*$. Thus, pooling firms obviously emit more and produce more compared to stand-alone trading. Equation (6) shows that after pooling marginal abatement costs are even lower than the permits price. Hence, also after pooling firms don't abate emissions in a cost-efficient manner. But since firms produce in a Cournot-duopoly too less output compared to the welfare-optimum, the higher output after pooling at least improves consumers' surplus.[13]

The most important result is that comparing expected profits with and without pooling, we can state that pooling always leads to lower expected profits, due to the assumption $0 < a < 1$.

$$E\{\Pi_i^S\}^* - E\{\Pi_i^P\}^{**} = \frac{p(4c(6(3-a)m + a(12+5a)p) + 27(3+6a+2a^2)pt)}{1296ct} > 0$$

Hence, there is no individual incentive for oligopolistic firms to pool their activities on the permits market. Pooling increases firms' output and, thus, decreases their returns. This negative effect always outweighs the positive effect of a lower expected fine.[14]

2.3 Collusive Pooling

In the section above, permits are not seen by the firms as a strategic instrument. But permits do not only cover emissions ex-post; they also can be interpreted, ex-ante, as a

[13] An explanation for the lower abatement costs is the very simplifying cost allocation rule $\phi_i = 1/2$.

[14] In case of a perfect product market, results differ. Hennessy and Roosen (1999) show that for perfect product market competition, if firms merge their activities on the permits market negative effects of uncertainty are reduced and firms' profits increase.

production input. Thus, permits are a strategic variable for an oligopolistic industry; the number of permits a firm is holding determines its output. Therefore, if both firms agree and commit to holding fewer permits, they produce less output. Since a lower production level shifts equilibrium from the competitive solution towards the cartel solution, firms' profits might increase.

When firms recognize this strategic interpretation of allowance holdings, they are interested in changing the natural sequence of actions: they aim to agree and commit on their emissions levels first and compete on the product market afterwards. The pooling option can be used to establish a forum for restricting emissions, which will probably be ignored by the regulator.

Thus, the timing of the game changes. At the first stage, firms cooperatively choose their amount of planned emissions. At the second stage, the trustee chooses the cost-minimizing amount of permits for the pool and firms play a Cournot-game at the product market. For the optimal decision at the first stage, they have to anticipate equilibrium at the second stage.[15]

Hence, at the first stage, firms maximize joint profits.

$$\begin{gathered}\textstyle\sum_{i\in\{1,2\}} E\{\Pi_i^P\}) = \\ P(Y)(y_1+y_2) - C(y_1,e_1) - C(y_2,e_2) - pQ + f\textstyle\int_{Q-e_1-e_2}^{a(e_1+e_2)}(e_1+e_2+Z-Q)dH(Z)\end{gathered}$$

Note that the cost allocation rule does not affect the equilibrium, as by definition $\phi_1 + \phi_2 = 1$.

At the second stage, firms choose their output simultaneously. Applying the necessary conditions (1), firm i's Cournot output is

$$y_i^{***}(e_i, e_{-i}) = \frac{4c^2e_i + 2cm + 4ce_it - 2ce_{-i}t + mt}{4c^2 + 8ct + 3t^2}$$

for all $i \in \{1, 2\}$.

At the second stage of the game, the trustee also chooses the total amount of permits that minimizes pool costs. The corresponding first order condition determining Q remains (4).

Applying the result of equation (4) and y_i^{***} to the expected profit function, the necessary condition for e_i, maximizing joint profits, is

$$0 = \begin{cases} -t(y_{-i}^{***}\frac{dy_i^{***}}{de_i} + y_i^{***}\frac{dy_{-i}^{***}}{de_i}) + 2c(y_i^{***} - e_i) - p - \frac{pa(3e_i^2 - e_{-i}^2)}{6e_i^2} \\ \text{if } e_i > e_{-i} \\ -t(y_{-i}^{***}\frac{dy_i^{***}}{de_i} + y_i^{***}\frac{dy_{-i}^{***}}{de_i}) + 2c(y_i^{***} - e_i) - p - \frac{pae_i}{3e_{-i}} \\ \text{if } e_i \leq e_{-i} \end{cases} \tag{7}$$

for all $i \in \{1, 2\}$.

[15] For model details and related literature we refer to Ehrhart et al. (2006).

Hence, the symmetric solution satisfying all necessary conditions is

$$y_i^{***} = \frac{(3m-(3+a)p)(2c+3t)}{3t(8c+9t)},$$

$$e_i^{***} = \frac{(3m-(3+a)p)(2c+3t)}{3t(8c+9t)} - \frac{2c(m+(3+a)p)+3(3+a)pt}{2c(8c+9t)}, \text{ and}$$

$$Q^{***} = \frac{2(3m-(3+a)p)(2c+3t)}{3t(8c+9t)} - \frac{2c(m+(3+a)p)+3(3+a)pt}{c(8c+9t)}$$

for all $i \in \{1,2\}$.

All second order conditions for y_i^{***}, e_i^{***}, and Q^{***} are satisfied. The corresponding Hesse-matrix is negative semidefinite. Thus, expected profits in equilibrium are

$$E\{\Pi_i^P\}^{***} =$$

$$\frac{4c^2(-3m+(3+a)p)^2+12c(3m^2-2(3+a)mp+(3+a)^2p^2)t+9(3+a)^2p^2t^2}{36ct(8c+9t)}.$$

As can easily be proved, firms cooperatively choosing emissions emit less and produce less than if firms pool in a non-cooperative way ($e_i^{***} < e_i^{**}$, $y_i^{***} < y_i^{**}$). Trivially, profits are higher if they coordinate their emissions. Therefore, firms have a clear incentive to coordinate their behavior on the permit market. Unfortunately, at the same time, the positive effects of pooling as described in Section 2.2 are counteracted.

Comparing the results of coordinated pooling with those of stand-alone trading, we can state that if $a < \frac{4c(m-p)}{4cp+3pt}$, firms produce less in case of coordinated pooling ($y^{***} < y^*$). The positive effects of pooling are lost, since consumers' surplus is even smaller in a coordinated pooling situation. If a is greater than the threshold, firms increase their production after coordinated pooling ($y^{***} > y^*$). Therefore, a positive effect of pooling on consumers' surplus exists.

The effect of coordinated pooling on emissions compared with stand-alone trading is also ambiguous. The amount of emissions again drops below the pre-pooling level ($e^{***} < e^*$) if $a < \frac{4c(m-p)}{4cp+3pt}$. Hence, it could be the case that the amount of emissions after coordinated pooling is even further away from the amount of emissions for a cost-efficient solution than if the firms were to act separately. If $a > \frac{4c(m-p)}{4cp+3pt}$ the amount increases above the level it had prior to pooling ($e^{***} > e^*$).

For both cases, profits after coordinated pooling can be higher than without pooling ($E\{\Pi_i^P\}^{***} > E\{\Pi_i^S\}^*$); hence, firms can actually have an incentive to pool.

The model above shows that pooling enables firms to jointly determine their emissions and thereby their output levels as well. These lower output levels can be interpreted as collusion between firms. Given this emissions restriction, the firms' behavior on the product market is still competitive and thus apparently does not arouse suspicion on the part of the regulator.

2.4 Extensions

Under the stated assumptions, our model implies that firms are only interested in pooling on the permits market when they can actually coordinate their emissions. Since such behavior mitigates or even eliminates the positive effects of pooling uncertainty, the regulator might prohibit allowances pools for firms from the same industry when market concentration exceeds a certain level.

Before making conclusions about the European emissions trading law, there are some refinements of our model we recommend to investigate. A crucial issue affecting for the result of Section 2.2 is what kind of cost allocation rule is used in the pool. Due to the assumption of identical firms, we assumed symmetric cost sharing. Obviously, there is no direct analogy if firms have different initial endowments or different emissions abatement costs, as is to be expected in practice. Moreover, if firms use an allocation rule depending on the initial endowment, the equilibrium also depends on the primary distribution of allowances.

Another interesting extension would to consider the effect of uncertain emissions on the overall permits trading system. If all firms face uncertain emissions, the firms' behavior also has an impact on the permits price.

Despite these extensions, our model presents the economic intuition, similar to Ehrhart et al. (2006), for the abuse of the pooling option for coordinating emissions and, thus, agreeing on production levels. Therefore, an originally welcome institution ultimately enables firms to behave in a collusive way – even when there is no verifiable evidence of collusion on the product market.

3 Legal Consequences

Although the model deviates in some parts from the EC emissions trading system, we use its results to analyze whether the pooling of activities on the permits market contravenes EC competition law. More precisely, the question is wether the "pooling option" conflicts with the European antitrust law, since it could lead to a restriction of production and therefore to a distortion of competition.

Article 81 of the EC Treaty addresses the impact of arrangements between undertakings on competition. Article 81(1) of the EC Treaty provides that "all agreements between undertakings ... which may affect trade between Member States and which have as their object or effect the prevention, restriction or distortion of competition" shall be prohibited.

The term 'undertaking' is defined by the European Court as any entity engaged in commercial activities regardless of its legal status. Therefore, firms participating in permits trading are clearly undertakings in the sense of Article 81 of the Treaty.[16]

Furthermore, in a trading pool as described in Article 28 of the Emission Trading Directive, firms trade their permits together. There is thus an agreement existing

[16] See Jones and Surfin (2004).

between the participants of the pool. They name a trustee who is responsible for trading permits and covering emissions on behalf of all participating firms. If firms abuse the pool as described in Section 2.3, firms also coordinate their emissions levels via the pool.

Since the participating undertakings in a pool are restricted to installations performing the same activity, the pool can be considered as a horizontal agreement. However, this kind of agreement does not have a negative impact on competition per se. In fact, this impact depends solely on the market structure and the strategies of the participants. Hence, any assessment of a permits trading pool must be done on a case-by-case basis.[17] In conclusion, a permits trading pool can be classified as a 'non-per-se' horizontal agreement between the participating undertakings.

For a pool to fall under Article 81(1), it must either have the objective or the effect of restricting competition and this restriction must be of an appreciable extent.

When the participating undertakings of the pool are competing on a perfect product market, there is no restriction of competition whatsoever; individual behavior has no impact on the market prices. Thus, a restriction of competition cannot be the objective of the firms. Rather, the effects of pooling are strictly positive under perfect competition. In terms of competition law it is more interesting to consider the case of oligopoly product markets. The model above shows that for firms in a symmetric duopoly industry, a pool always provides an incentive to agree on emissions, with the objective of limiting and controlling production. Under these circumstances, the main objective is a restriction of competition on the product market. It remains to be checked, whether the agreement has an appreciable potential impact on competition. The commission states that if the undertakings represent more than 5 per cent of the total production of the common market, an agreement is likely to be appreciable and Article 81(1) is likely to apply.[18] Since we have assumed a duopoly product market, this condition is satisfied.

If it is not the objective of the agreement to restrict competition, then it must still be investigated, whether it has the effect of causing appreciably less competition. In this case, Article 81(1) also applies. However, since the results of Section 2 show that there are only negative effects on the product market if firms agree cooperatively on their emissions, and hence already have the objective to restrict competition this case is not relevant in our framework.

In addition to the distortion of competition, Article 81(1) also requires that there may be an effect on trade between Member States. This is the case when the agreement has or will have an influence on the pattern of trade between Member States. Note, however, that this effect does not necessarily have to be harmful or negative; it is not even necessary that trade is actually affected at all. It is sufficient to show that such an effect might be possible. In our scenario, this is the case even if firms' output is only traded within one Member State. If a national agreement, restricting competition, dominates the whole or a large part of the market, it is likely to rein-

[17] See Faull and Nikpay (1999).

[18] See Faull and Nikpay (1999).

force compartmentalization of the market on a national basis. Hence, penetration by undertakings from other Member States becomes more difficult.[19] Another approach for showing possible effects on trade is to consider if the market structure changes in such a way that firms shut down. This is rather unlikely to be achieved via the abuse of the pooling option. All firms in the oligopoly market profit from an increase in the product price, even if they are not participating in the pool.[20]

If Article 81(1) must be applied to a permit trading pool, it is, pursuant to Article 81(2), automatically void. The consequences of Article 81(2) are not a matter of European law and are determined by national courts.

If an agreement is prohibited by Article 81(1) of the Treaty, it must be determined if it can be salvaged by Article 81(3), which states that such an agreement is permitted if it has a positive effect on welfare and consumers also benefit. To test if an exemption can be granted, we distinguish the two cases from Section 2.3 as follows: Firstly, the industry output level drops after pooling below the level it had before pooling. Secondly, firms abuse the pool to agree on their emissions, but output increases above the production level before pooling.

In the first case, the effect of a permit trading pool is strictly negative for the consumers and there are no benefits from a pool. Only firms benefit from increased profits. Therefore, Article 81(3) cannot be applied.

The second case is slightly more complicated. Here all participants benefit. The firms benefit from less uncertainty on non-compliance as well as from agreeing on their emissions. Consumers also benefit from the reduced uncertainty, since the product price decreases with an increased output. Thus, the condition that consumers benefit in some way is satisfied. But Article 81(3) only applies if there is no restriction on competition, which is not absolutely necessary for the positive benefits. In our case, the positive effect for the consumer only arises from the reduced uncertainty after pooling. The coordination of the emissions within the pool always reduces the positive effects of pooling and is therefore counteractive for the benefit. Thus, for achieving the benefit, only a pool in which firms do not agree to lower emissions has no unnecessary restrictions on competition. But as the model above shows, firms will always try to coordinate themselves within a pool in order to increase profits. Thus, even if consumers' surplus increases after pooling, it can be expected that firms behave collusively, and hence no exemption due to Article 81(3) can be granted.

This short legal analysis shows that pooling permits activities as defined in Article 28 of the Emission Trading Directive can conflict European competition law. If and how mainly depends on the market structure of the product market. Thus, a general prohibition of pooling is not justified. But antitrust authorities may be well advised to scrutinize pools within an oligopolistic industry very strictly.

[19] See Judgement of 17 October 1972 in Case 8/72 , Cementhandelsaren v. Commision [1972], ECR 977, or Case 42/84, Remia v. Commision [1985], ECR 2545, para 22.

[20] Even if it is not possible to foresee any influence on trade between Member States, permits trading pools may nevertheless still violate national competition laws. This would be the case with respect to the German Competition Law (GWB), according to its first paragraph.

4 Conclusion

This paper presented an analysis of the impact of stochastic emissions on allowances trading in oligopolistic product markets as well as the ambiguous effects trading pools have on firms behavior. Uncertain emissions lead to less investing in permits and more abatement compared to a cost-efficient permits trading system. On the product market, uncertain emissions cause a reduction in output, which diminishes consumers' surplus. Pooling activities on the permits market mitigates undesirable randomness of emissions, resulting in increasing emissions and output. However, the model has also shown that firms may have no incentive to pool since doing so reduces profits. In case firms interpret permits as a strategic input factor, we demonstrate that they are interested in committing on their allowances holdings within the pool before competing on the product market. Firms that coordinate their permits holdings agree to buy fewer permits than if they pool uncoordinatedly. This reduction always decreases output and hence, effects consumers' surplus negatively, and firms' profits positively. In the worst scenario, in which firms recognize the strategic component of permits and exploit it, they emit even less than without pooling and their output drops below the level before pooling. Recapitulating, pooling can enable an oligopolistic industry to collude on its output level via the permits market in a non-obvious manner.

A legal analysis has shown that, when firms abuse the pooling-option to coordinate their emissions, this agreement on the allowances market constitutes a violation of EC competition law. Even if coordinated pooling increases output above the output level without pooling, EC antitrust law prohibit such pools. However, a general prohibition of pooling is not enforceable by EC competition law, since under perfect market conditions pooling has a strictly positive impact on allowances and product markets. Therefore, the results presented in this paper suggest that the European Commission (or National Competition Authorities) should observe pools within oligopolies very carefully.

References

Ben-David, S., D. Brookshire, S. Burness, M. McKee, and C. Schmidt (2000): "Attitudes Towards Risk and Compliance in Emission Permit Markets," *Land Economics*, 76(4), pp. 590–600.

Betz, R., W. Eichhammer, and J. Schleich (2004): "Designing National Allocation Plans for EU emissions trading - A First Analysis of the Outcome," *Energy & Environment*, 15(3), pp. 375–425.

Betz, R., J. Schleich, and C. Wartmann (2003): *Flexible Instrumente im Klimaschutz*, Ministerium für Umwelt und Verkehr Baden-Württemberg, Stuttgart.

Carlson, D. A. and A. M. Sholtz (1994): "Designing Pollution Market Instruments: Cases of Uncertainty," *Contemporary Economic Policy*, 12(4), pp. 114–125.

Ehrhart, K.-M., C. Hoppe, and R. Löschel (2006): "Abuse of EU Emissions Trading for Tacit Collusion," Universität Karlsruhe (TH).

Faull, J. and A. Nikpay (1999): *The EC Law of Competition*, Oxford University Press, Oxford.

Graichen, P. and T. Requate (2005): "Der steinige Weg von der Theorie in die Praxis des Emissionshandels: Die EU-Richtlinie zum CO2-Emissionshandel und ihre nationale Umsetzung." *Perspektiven der Wirtschaftspolitik*, 6(1), pp. 41–56.

Hennessy, D. A. and J. Roosen (1999): "Stochastic pollution, permits, and merger incentives," *Journal of Environmental Economics and Management*, 37, pp. 211–232.

Jones, A. and B. Surfin (2004): *EC Competition Law*, 2. edition, Oxford University Press, Oxford.

Klinski, S. (2003): "Emissionshandel: Ganz, zum Teil oder in Pools," *Technik & Management*, 03/2, pp. 64–65.

Maeda, A. and T. Tezuka (2004): "Intertemporal Trading Strategy under Emissions Uncertainty," Kyoto University.

Montero, J.-P. (1997): "Marketable pollution permits with uncertainty and transaction costs," *Resources and Energy Economics*, 20(1), pp. 27–50.

Montgomery, W. D. (1972): "Markets in licenses and efficient pollution control programs," *Journal of Economic Theory*, 5(3), pp. 395–418.

Mrozek, J. R. and A. G. Keeler (2004): "Pooling of uncertainty: enforcing tradable permits regulation when emissions are stochastic," *Environmental resource economics*, 29(4), pp. 459–481.

Newell, R. G. and W. A. Pizer (2000): "Regulating Stock Externalities Under Uncertainty," `http://www.rff.org/rff/Documents/RFF-DP-99-10-REV.pdf`, resources for the Future, Discussion Paper 99-10-REV.

Spieth, W. and C. Röder-Persson (2003): "Umsetzung der Emissionshandels-Richtlinie in Deutschland," *Energiewirtschaftliche Tagesfragen*, 53(6), pp. 390–394.

Trust and the Law

Christoph Sorge[1], Thomas Dreier[2], and Martina Zitterbart[1]

[1] Institute of Telematics,
Universität Karlsruhe (TH)
{sorge,zit}@tm.uka.de

[2] Institute of Information Law,
Universität Karlsruhe (TH)
dreier@ira.uka.de

Summary. In a world full of risk, trust plays an important role: most transactions performed in electronic commerce could not take place without trust in the transaction partners. Law, too, is a prerequisite for all kinds of transactions. But how do these concepts relate to each other? How does this relationship change in case of transactions over computer networks?

At first glance, trust and law seem to be orthogonal concepts. However, we will show that trust plays an important role in law—and law is an important factor supporting trust, on the other hand.

This article makes a case for the need of the concept of trust in legal research by pointing out how trust and law relate to each other.

In recent years, computer science has also started to address the topic of trust. We will focus on the communication of trust relationships via technological means in the context of electronic commerce. This in turn leads to new legal considerations, which are discussed in the final part of the article.

1 Introduction

The title of this article may have prompted you to wonder if there is indeed a relationship between trust and law, and what, if anything, the two concepts have to do with the topic of "information management and marketing engineering". After all, the Graduate School publishing this book is dedicated to these very disciplines.

Indeed, we will show that numerous disciplines are concerned with the topic of trust. Law is one of them. This is not only true for legal philosophy, but also for criminal law and civil law. In fact, the law has recently been confronted with a new trust-related challenge: with the growing success of electronic commerce (B2C and C2C), new trust relationships between transaction partners have to be built up. Reputation systems have grown popular as one way of establishing these relationships. As buying decisions are significantly influenced by these systems, there is an incentive to tamper with them, entailing the falsification of the information provided. The law is therefore charged with reacting to potential manipulations while simultaneously enabling the legitimate use of reputation systems and protecting users' freedom of opinion (the right to express one's opinion, protected by the German constitution and closely related to the freedom of speech).

Similarly, trust plays an important role in market engineering. Particularly for transactions in electronic markets, trust is crucial. It is often difficult or too costly to enforce the rendering of a service or the making of a payment. Any designer of a market must take this issue into consideration—a way must be found to instill trust among market participants as well in the market platform itself.

In this article, we focus on the legal aspects of trust, as they also have an important impact on the work of market designers.

2 Terminology

Three terms are particularly important for the understanding of this article.

- The first term is *trust*.

  What is trust? The term has been defined in numerous ways. According to Sztompka (2001), "trust is a bet about the future contingent actions of others". "Others" are not necessarily natural persons, but can also be companies, the environment or other objects. "Bet" refers to the risk always incurred when you trust someone. The term "future" may be replaced by "imperceivable". If you trust someone, you assume (or bet) that that his behavior will conform to your expectations—now or in the future.

  Note that trust is *not* security. "Secure" can be defined as "free from risk". In a perfectly secure world, only one outcome is possible. There is no need for "bets". Trust is only required in insecure or uncertain environments. Given the fact that absolute security does not exist, the concept of trust is far from obsolete.
- The second term is *reputation*. We define an entity's reputation as its trustworthiness, as perceived by a group of other entities. Accordingly, *reputation systems* are systems which provide information to enable the assessment of an entity's trustworthiness or reputation.
- The last term is *law*. Law has been defined as "the discipline and profession concerned with the customs, practices, and rules of conduct of a community that are recognized as binding by the community." (Encyclopaedia Brittanica, 2005, entry "law").

3 Relationship of Trust and Law

According to Darmstaedter (1948), law aims at making trust dispensable. By imposing norms and sanctions to redress violation of these norms, law was meant to improve a trusting person's position: In the event that their trust was exploited or misplaced, the law was there to protect them. However, the legal system could only redress a lack of willingness, and not a lack of ability, to render a service.

We can only share this view to a limited extent. Indeed, the legal system affects the will rather than the ability to render a service. But trust can never be fully replaced by law: There is always a risk that the trustee will circumvent the legal system or exploit trust in non-sanctioned or unforeseeable ways.

Thus, the legal system can promote trust by reducing the risk involved—but cannot replace it. Instead, it is a precondition for trust. In many cases, the legal system aids in the realization of the economic advantages brought about by trust.

With respect to Italian law, Memmo et al. (2003) have discussed the law's supporting role with respect to trust. The authors differentiate between certain kinds of trust:

- *Situation trust*, or the trust in the existence of "certain legally relevant situations".
- *Behavior trust*, or the trust in the behavior of an agent.

Table 1: Examples of the protection of trust in different fields of law

	Situation trust	**Behavior trust**
Criminal Law	Error as to the prohibited nature of an act (§ 17 StGB)[1]	Protection of confidential information (e.g. §§ 201, 203 StGB)
Civil Law: Direct Protection	Public registers (e.g. commercial register)	Fidelity liability
Law of torts	—	Compensation claims (§§ 823, 826 BGB)

Table 1 provides some examples of measures designed to protect both forms of trust in different fields of law. Our focus will be on the second form of trust (behavior trust), however. Three aspects strike us as particularly important, based on the example of the German legal system:

- Trust that is based on laws influencing other parties' behavior. Such laws can be found in
  - Criminal law. For example, you entrust your doctor with confidential health information because you trust that this information will not be revealed without your consent. One reason for this is that your doctor can be punished for revealing information (§ 203 I of the German penal code StGB).
  - Civil law, especially the law of torts. Consider, for example, § 823 II of the German civil code (BGB): If there is a law that serves your protection, you can

[1] German law protects trust in the lawfulness of an act even if the act was in fact unlawful. However, this protection is limited to inevitable errors.

trust that others will abide by this law. If they do not, they will have to pay compensation. This eventuality serves to deter the other party from breaking the law in the first place; however, if it does not, you will be compensated for damages.

- We call the second aspect “direct protection of trust by the law”. We give an overview of this form of protection in Section 4.
- The third aspect is the legal treatment of trust-supporting mechanisms outside the legal system. The most prominent example is reputation systems, which are described in Section 5.

4 Direct Protection of Behavior Trust by the Law

There are situations in which a person’s trust in another person’s behavior is directly protected by the law. This means that, if person A trusts person B to behave in a certain way, and B behaves differently, the law dictates a (partial or complete) compensation of A’s damages. It is worth mentioning that this compensation is awarded to redress the violation of A’s trust, and not to communicate disapproval of B’s actions per se on the part of the legal system. If compensation is to be paid by B, we use the term *fidelity liability* (Vertrauenshaftung).

Several legal regulations deal with fidelity liability; we list the most important ones here, based on German legislation (cmp. Schäfer and Ott, 2005, pp. 525–528):

- *The law of representation.* §§ 170, 171 II, 172 II, 173 and 179 I BGB protect a person’s trust in another person’s authorization as a representative: In case of §§ 170, 171 II and 172 II BGB, the authorization remains valid if it was announced to the contractual partner in certain ways (with an exception given in §173 BGB); if the contractual partner trusts an unauthorized representative, he is protected by § 179 I BGB.
- *Liability for a declaration.* If a person rescinds a declaration of intention, he has to pay a compensation for the damage caused due to the recipient’s trust in the validity of the declaration (§ 122 BGB). § 179 BGB also constitutes a liability for a declaration.
- *Liability for creating the appearance of a legal position.* If someone creates the semblance of a certain legal situation in an attributable fashion, and someone else makes arrangements based upon his trust in the existence of this legal situation, the trusting person can appeal to that situation (Canaris, 1971, pp. 526–528). Again, the law of representation can serve as an example: If a person creates the semblance of another person being authorized to represent him, he can be bound by declarations of this person, even if there was in fact no such authorization.
- *Liability for contradictory behavior.* This concept is based on § 242 BGB: If someone trusts in another person’s behavior, and this trust is worthy of protection, the trusting person is protected if the trustee contradicts his previous behavior (Schäfer and Ott, 2005, p. 528).

But why does the law protect trust in these cases? From an economic perspective, Schäfer and Ott (2005, p. 517) have identified four requirements for the legal protection of trust:

- *The asymmetry of information costs.* If it is cheaper for one party to obtain a piece of information, it may make sense to protect the other party relying on this information. This is particularly true if the information concerns the first party's future behavior.
- *The societal productivity of information.* If, considering society as a whole, the costs of obtaining the information are lower than the benefit from the information, it is better to obtain the information. Therefore, a party's trust in the other party obtaining the information should be protected in this case. For example, information pertaining to whether someone is authorized as a representative can be valuable: The damage caused if someone relies on the validity of a contract can be greater than the benefit for the person who does not provide information about the authorization. That is why the contractual partner's trust is protected.
- *The existence of a trust premium.* The trusted party, being held liable, must receive compensation for the costs incurred from the protection of trust. In a market, this is normally the case; however, if the trusted party just provides the information as a favor and free of charge, protecting the other party's trust in this information is inappropriate. For a more exhaustive discussion of trust premiums, see Schäfer and Ott (2005, pp. 519–522).
- *The relation of trust premium and opportunism premium.* If the benefit from exploiting another party's trust (opportunism premium) is higher than the premium for a trustworthy behavior (trust premium), this is an argument for the legal protection of trust; if, on the other hand, the opportunism premium is lower than the trust premium, there is no incentive to exploit trust, and legal protection might become unnecessary. In many markets, sanctions for opportunistic behavior result from a negative impact on the opportunistic party's reputation; improving the quality of this mechanism can therefore make legal protection unnecessary. Reputation systems, which are considered in the next section, are an approach to achieving this goal.

5 Law and the Communication of Trust

Before addressing legal issues surrounding reputation systems, we begin with a short introduction to the systems themselves and a discussion of the possible pitfalls.

5.1 Reputation and Reputation Systems

As defined above, a reputation system's main purpose is to distribute information about the (perceived) trustworthiness of certain entities. Even long before the Internet existed, the same task had to be performed. We therefore commence our discussion with a description of classical reputation mechanisms.

Classical reputation mechanisms.

Let us consider a simple example of a decision where trust plays a role: You want to buy meat. Even nowadays, you cannot always be sure about the contents of a sausage, or you might realize too late that your purchase was rotten. Law can help in this situation: If a certain quality standard is not observed, and your butcher did not inform you about the lower quality prior to the sale, you may be entitled to warranty and compensation claims. However, this only helps you ex post, and enforcing a claim may be time-consuming. Therefore, you probably want to avoid this situation in the first place. So how do you select a butcher? Obvious factors in making your decision might be the location, the visual impression of the shop, and so on. However, these aspects are not our main focus.

Instead, consider the reputation-based approach: You ask someone (preferably someone you trust) about the quality of a certain butcher. In the simplest case, this person knows the butcher and can give you an answer from his or her own first-hand experience. Alternatively, the person may know (and trust) other people who happen to be customers of this particular butcher—or who in turn know (and trust) other people, and so forth. Of course, you can also ask more than one person yourself. If you receive different answers, you can decide which answer is most likely correct. You might choose the answer of the person you believe to be most trustworthy, or (even better) you may weight the answers:

- according to your estimation of the person's trustworthiness,
- according to the person's sources (first-hand experience, friends, rumors, ...).

At this point, we feel it is no longer appropriate to say that you "trust someone". Maybe you do not even trust your best friend's opinion on butchers, because he is a vegetarian. It seems necessary to differentiate trust in certain domains or contexts. In other words: you do not just trust someone, but you trust someone with respect to a domain.

In this way, a multitude of trust networks can be formed: In any given domain, people communicate their trust assessments (or, speaking more generally, their evaluations)—and change them, depending on other people's opinions. Reputation systems try to adopt that principle and apply it to computer networks.

Reputation systems.

Various kinds of reputation systems are currently in operation. As the most important classification, we differentiate between centralized and distributed reputation systems. In a centralized reputation system (such as the eBay reputation system; Resnick and Zeckhauser, 2002), a central server manages all trust assessments. Clients merely add assessments to the system and query for other entities' reputations. In a distributed system, however, there is no central entity. Several nodes have to cooperate in order to compute a node's reputation value. Note that there can still be a centralized trust model in a distributed system, and vice versa.

Other classification criteria include:

- the representation of evaluations (e.g. as a text or a number).
- the point in time at which trust assessments can be made (anytime or only a within a certain period after a transaction has taken place).
- the trust model used. Is there a single trust anchor, or are there just trust relationships between the participants (web-of-trust)? If a system computes a global (system-wide) reputation value for each node, we can think of "the system" as a single trust anchor.
- the domains for which the reputation system is designed.
- the subjectivity of evaluations. Does an evaluation consist of facts (for example, "The user delivered my book within two days") or subjective assessments (for example, "I trust this person very much")?

We will see that these classification criteria also have an impact on the legal issues related to reputation systems.

5.2 Legal Consideration for Reputation Systems

Lawfulness of evaluations.

A user mainly interacts with a reputation system when:

- reading evaluations or
- writing evaluations.

While there are no legal consequences pertaining to the act of reading the evaluations, the contents of user-submitted evaluations may be subject to legal scrutiny.

A number of regulations stipulate entitlement to a claim for restraint or compensation, the most important ones being:

- §§ 8 I, 9, 3 UWG (Gesetz gegen den unlauteren Wettbewerb, law against unfair competition). If a competitor submits an unjustified negative evaluation, this impairs competition to the evaluated party's disadvantage. If the impairment is not insubstantial, this leads to claims for both restraint (§ 8 I UWG) and compensation (§ 9 UWG). Schmitz and Laun (2005, p. 208) come to a different conclusion; typically, the aim of an evaluation was not to increase sales of one's own or a third party's products. However, an evaluation can easily be used to harm a competitor's position on a market; therefore, we believe this claim to be relevant.
- § 824 BGB. If someone spreads inaccurate statements of fact that endanger another person's credit, he or she has to pay compensation if he or she knew (or had to know) that the information was incorrect.
- §§ 1004, 823 I BGB. § 823 I BGB protects a number of absolute rights; in case of violation of any one of these rights, the violator has to pay compensation. In case of reputation systems, an "other right" might be concerned: the right to conduct an established and carried on business. Negative evaluations can considerably harm such a business. Note, however, that the scope of this right is quite limited: companies are not supposed to be privileged over private persons (von Staudinger, 1999,

§ 823, recital D2). Furthermore, the right is only subsidiary. It fills gaps in the written law, but only where such gaps exist (von Staudinger, 1999, § 823, recital D20). Where e.g. competition law applies, this is not the case.
There is also a controversy on whether the right to the established and carried on business is an "other right" in the sense of § 823 I BGB; however, this controversy is of limited practical relevance (von Staudinger, 1999, § 823, recital D3).

- § 826 BGB. If someone intentionally harms another person contrary to public policy, the other person is entitled to compensation. However, this norm is rarely applicable in case of reputation systems (Schmitz and Laun, 2005, p. 208).

In all three cases, however, the interpretation of the norms requires a consideration of the legal interests protected by German constitutional law:

- the evaluator's right to freedom of opinion[2]—allowing the evaluator to express his opinion e.g. by writing evaluations—on the one hand, and
- the right to conduct an established and carried on business (or the general personality right in case of a private individual) of the evaluee—which might be harmed by the evaluation—on the other hand.

An opinion, as protected by the German constitution, has been defined as a statement that is "characterized by elements of comment, of points of view or of meaning".[3] This means that a subjective view on a person is protected; this is not true for pure statements of fact. The differentiation, however, is often difficult: While an inaccurate statement of fact is not protected *as such* (BVerfG, 1980, p. 2073), it is protected by article 5 I GG when it is part of a statement of opinion. It must be determined on a case-by-case basis whether a statement is primarily factual in nature, or whether a fact has just been used to support the formation of an opinion (cmp. BVerfG (1983))—in this case, we would speak of opinion-relevant statements of fact (Nolte and Tams, 2004, p. 111).

Naturally, freedom of opinion is not unlimited. Even a statement falling within the protective scope of article 5 I can be illegal. Article 5 II GG states that the freedom of opinion finds its "limits in the provisions of general laws [...] and in the right to personal honor." That means that a balancing of the basic rights of evaluating and evaluated person is usually necessary. In the case of vilifying criticism ("Schmähkritik" in German), the balance is tipped in favor of the individual subjected to the criticism (BVerfG, 2001). Criticism can be said to be vilifying if it serves to "defame, vilify or impair the individual concerned" (Schmitz and Laun, 2005, p. 209), while the alleged matter takes a back seat.

If the evaluee has voluntarily subjected him- or herself to evaluations, vilifying criticism also constitutes the only instance in which an evaluation becomes illegal: If someone offers goods or services or even registers in a reputation system voluntarily, this person benefits from the possibility of evaluating others and being evaluated. Therefore, he or she is only worthy of protection to a very limited extent.

[2] article 5 I of the German constitution (Grundgesetz, GG)
[3] "durch die Elemente der Stellungnahme, des Dafürhaltens oder Meinens geprägt" (BVerfG, 1983)

While textual evaluations are open for interpretation as to whether or not they constitute vilifying criticism, it is very difficult to assess numerical evaluations in terms of vilifying content due to their abstract (and therefore non-subjective) nature. The authors of this article are not aware of any court decision dealing with a "negative" numerical evaluation without additional text. In most cases, it is hard to prove malicious intent on the part of a numerical evaluation. Exceptions may occur if the evaluator infringes the rules governing the evaluation process, or if this person is a competitor of the evaluee.

More important than the scale used is the domain of evaluation. The more subjective the evaluations in the respective domain are, the less can they be controlled by a court. If participants evaluate the promptness of a transaction settlement, there is room for an objective check. This will typically be the case if merchants trading in standardized goods are evaluated. When buying a book online, a customer will probably not judge a store based on the quality of the book (i.e. in terms of style or content, which is a rather subjective assessment), but more likely on the speed of delivery, availability of the shop systems, and other measurable attributes, including perhaps the physical condition of the book.

However, if an evaluation states to what extent the evaluator is willing to entrust personal information to the evaluee—a highly personal decision—, he or she can do so without any rational foundation. Redress by a judge is impossible in such a case.

Yet technology itself can partly solve problems that law cannot: For example, filtering out evaluations performed by persons who have proven themselves to be untrustworthy significantly reduces the damage they can cause.

Liability of platform operators.

If someone suffers damages from an illegal evaluation, it may be in this person's interest to raise a claim against the platform operator (if there is any) instead of (or in addition to) the evaluator: the chance of the platform operator being able to pay compensation is higher than in the case of the evaluator.

But is there such a claim?

Usually, reputation systems are teleservices; therefore, their responsibilities are defined in the body of law on teleservices (Teledienstegesetz, TDG). According to § 8 I TDG, service providers are responsible for the accuracy of their *own* information (in conformity with the relevant general laws). This is also true for any information that they appropriate as their own.

Two arguments have been advanced for the purpose of assigning liability to platform owners with respect to user evaluations based on the appropriation of these evaluations. Both arguments have been (rightly) rejected by Schmitz and Laun (2005, p. 21):

- Platform providers routinely ask their customers to grant usage rights to their evaluations. However, this is done purely for copyright reasons.

- The platform has an economic interest in the evaluations. However, the same can be said for access providers—yet, they are not responsible for the information transmitted.

As a consequence, evaluations cannot be considered as a platform operator's own information.

If the operator aggregates evaluations (in order to give its customers a summary of the evaluators' assessments), however, the situation changes. Though aggregation is a reduction of information, a user of the system will usually consider the information it generates to be new—and this is the important aspect from a legal perspective. The aggregation function used always contains a valuation by the operator: Depending on how single evaluations are weighted, the result can form a completely different image of the evaluated person. For example, evaluations can be weighted:

- according to their age; greater weight is assigned to the most recent evaluations.
- according to the transaction value in case of transaction-based reputation systems.
- according to the evaluator's reputation.

A platform operator enjoys broad discretion with respect to choosing an aggregation algorithm. Therefore, any information generated by it must be considered his own information.

This means that the platform provider is responsible for the aggregated evaluations according to the general laws, insofar as new information was generated by the aggregation algorithm.

But what about other information (i.e. the users' evaluations)?

§ 11 TDG deals with the responsibility for information that a service provider stores "for a user". At first glance, the fact that other users can access the evaluation might argue against applying § 11 TDG to reputation systems. However, the situation is similar in case of hosting providers—and the statement of grounds of § 11 TDG (Bundestag, 2001) explicitly mentions its applicability to hosting providers. Yet the user submitting an evaluation does not even have an interest in storing it (in contrast to other users of the system, who base their decisions on such an evaluation). As the storing is triggered by the evaluator, we can, however, assume that the information is stored for him. This is in accord with the legislature's intention (cmp. Schmitz and Laun (2005, p. 211)).

Therefore, according to § 11 TDG, the operator of a reputation system is not responsible for evaluations as long as he is not aware of illegal contents, or deletes these contents (or inhibits access to them) without undue delay after coming to know about them. Note that awareness of illegal contents is required—not awareness of the illegality of contents.

Having discussed the impact of law on reputation systems, we go on to address the impact of reputation systems on law in the next section.

5.3 Reputation as a Substitute for Law?

In Section 3, we came to the conclusion that law cannot make trust dispensable. Instead, law promotes trust. However, the communication of trust in reputation systems can be a partial substitute for law.

Consider the perspective of a customer selecting a merchant: The availability of information about possible contractual partners influences his choice, increasing the probability that he will choose a trustworthy merchant and the transaction goes well. This means that there will be less need to fall back on the legal system.

From the merchant's perspective, the existence of an effective reputation mechanism means that opportunistic behavior is less likely to pay: it will result in a negative evaluation. Due to the reputation system, other customers will be likely to choose different merchants, reducing the opportunistic party's future revenues. The opportunism premium is reduced and is likely to become negative. To put it in the words of Schäfer and Ott (2005, p. 522), the market can protect trust more cheaply and effectively than the courts or the state.

Yet, effective reputation systems do not exist on all markets. Even if they did, they might only provide an incentive to behave in a trustworthy manner, but they would not guarantee trustworthy behavior one hundred per cent. Individuals could still be harmed—for example, a merchant may simply act irrationally. Therefore, legal instruments are still required—not just for the regulation of the reputation systems themselves, but also for cases in which they fail. Reputation can, however, serve as a mechanism to reduce the likelihood of legal disputes.

6 Conclusion

In this article, we investigated the relationship of trust and ´the law. We have shown that the law promotes trust rather than substituting for it. Examples include criminal law and the law of torts, as they reduce the chance of trust being exploited—at least in certain fields.

Reputation systems are an example of support for trust by means of technology. But they too require a legal framework. We have outlined this framework, showing that it is often difficult to balance the interests of the evaluator against those of the evaluee. However, more advanced reputation systems are able to solve or at least reduce problems that the law cannot always.

References

Bundestag (2001): "Entwurf eines Gesetzes über rechtliche Rahmenbedingungen für den elektronischen Geschäftsverkehr," Deutscher Bundestag, Drucksache 14/6098.

BVerfG (1980): "Verhältnis des allgemeinen Persönlichkeitsrechts zur Meinungsfreiheit," *Neue Juristische Wochenschrift*, 33(38), pp. 2072–2073, BVerfG, Beschluß vom 3. 6. 1980 – 1 BvR 797/78.

BVerfG (1983): "Bezeichnung der CSU als "NPD Europas" im Wahlkampf," *Neue Juristische Wochenschrift*, 36(25), pp. 1415–1417, BVerfG, Beschluß vom 22.06.1982 – 1 BvR 1376/79.

BVerfG (2001): "Entscheidung des Bundesverfassungsgerichts vom 1.8.2001," Az. 1 BvR 1906/97.

Canaris, C.-W. (1971): *Die Vertrauenshaftung im deutschen Privatrecht*, *Münchener Universitätsschriften: Reihe der Juristischen Fakultät*, vol. 16, C.H. Beck'sche Verlagsbuchhandlung, München.

Darmstaedter, F. (1948): "Recht und Jurist," *Süddeutsche Juristenzeitung*, 3, pp. 430–436.

Encyclopaedia Brittanica (2005): "Encyclopaedia Brittanica online," `http://www.britannica.com/`.

Memmo, D., G. Sartor, and G. Q. di Cardano (2003): "Trust, Reliance, Good Faith, and the Law," in: Paddy Nixon and Sotirios Terzis (ed.), *Proceedings of the First International Conference on Trust Management (iTrust 2003)*, LNCS 2692, Heraklion, Crete, Grece, pp. 150–164, `http://link.springer.de/link/service/series/0558/tocs/t2692.htm`.

Nolte, M. and C. J. Tams (2004): "Grundfälle zu Art. 5 I 1 GG," *Juristische Schulung*, 44(2), pp. 111–113.

Resnick, P. and R. Zeckhauser (2002): "Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System," in: M. Baye (ed.), *The Economics of The Internet and E-Commerce*, *Advances in Applied Microeconomics*, vol. 11, Elsevier.

Schäfer, H.-B. and C. Ott (2005): *Lehrbuch der ökonomischen Analyse des Zivilrechts*, 4th edition, Springer, Berlin, Heidelberg.

Schmitz, F. and S. Laun (2005): "Die Haftung kommerzieller Meinungsportale im Internet," *Multimedia und Recht*, 8(4), pp. 208–213.

Sztompka, P. (2001): *International Encyclopedia of the Social and Behavioral Sciences*, vol. 23, chapter Trust: Cultural Concerns, Elsevier, Amsterdam, Paris, New York, pp. 15913–15917.

von Staudinger, J. (1999): *Kommentar zum Bürgerlichen Gesetzbuch mit Einführungsgesetz und Nebengesetzen*, vol. §§ 823-825, Sellier-de Gruyter, bearbeitet von Johannes Hager.

Legal Promotion of Open Access Archives and Possible Implications: The Proposal of the German Conference of Education Ministers

Kendra Stockmar[1], Thomas Dreier[1], and Andreas Geyer-Schulz[2]

[1] Institute of Information Law,
Universität Karlsruhe (TH)
{stockmar,dreier}@ira.uka.de

[2] Institute of Information Systems and Management,
Universität Karlsruhe (TH)
andreas.geyer-schulz@em.uni-karlsruhe.de

Summary. To promote access to scientific publications through institutional archives, the German Conference of Education Ministers (Kultusministerkonferenz) has proposed a legal mandate to grant the universities a non-exclusive right of use, also referred to as "second publication right" (Zweitveröffentlichungsrecht), to the scientific works produced by their employees. This proposal raises some questions concerning the interpretation of the wording and the corresponding legal consequences. Additionally, due to the proposal's very real potential to affect academic freedom, a discussion about its constitutional legitimacy is currently taking place. A close examination of its wording reveals a number of legal ambiguities. In this paper, we address the potential ramifications of legal change in this area. We also provide a brief introduction to the constitutional legitimacy problem inherent in the proposal.

1 Introduction

Scientific communication, which includes the organization and principles of scientific publishing, has been a frequent topic of discussion in recent times. In particular, the so-called "journal crisis", which pertains to rising journal costs and insufficient library budgets, has generated some new ideas for enhancing access to research results or providing what is also known as "open access." One current approach to providing open access to scientific data entails placing peer-reviewed journal articles into archives that are openly available via the Internet. These archives can be institutional, discipline-based and/or centralized.

Recently, there have been various attempts to promote the population of these archives. One such attempt proposes legally mandating researchers at universities to grant non-exclusive usage rights to their works to their institutions, thereby enabling the placement of these works into open access archives. In this regard, the Conference of German Education Ministers (KMK) has issued a proposal for changing the copyright law pertaining to works produced by university employees (Kultusministerkonferenz, 2005). The proposal has generated mixed reviews. Mittler (2005) supports the idea in principle while Hansen (2005) remains more sceptical. In any case, if the proposal comes into force in its current incarnation, several conflicts threaten to arise. There are also some aspects in the proposal which are vague and require clarification before a legal change of this magnitude comes into force. Otherwise, legal uncertainties and unintended consequences could result. The goal of this paper will be to show the possible implications of this proposal and work out the open questions associated with it.

2 Wording and Background of the KMK-proposal for a new § 43 II UrhG (German Copyright Law)

Against the background of the open access movement, supported among others by the signatories of the Berlin Declaration (2003), the German educational ministers came up with the following proposal for a legal change in the framework of the copyright law reform.[1] This second clause is to be added to the existing statute (§ 43 UrhG) (Kultusministerkonferenz, 2005, p. 104 sq.):

§ 43 clause 2 UrhG/KMK-Proposal

- *Universities and research establishments are entitled to a non-exclusive usage right to the scientific works of their employees which have been produced within the scope of their teaching and research activities and which are intended to be published.*
- *The employees are required to announce the intention of publication without delay to the university or research establishment and to provide these in digital format.*

The intention behind this proposal is to make publicly funded research results available through open access archives. If universities are granted non-exclusive usage rights, it should theoretically be possible to place every article falling into this category into archives that are accessible online. According to the presentation of the proposal, the overall goal is to promote further access to scientific literature to counteract one aspect of the publication crisis, i.e. the cancellation of subscriptions by some libraries. By having these archives, the literature would at least be accessible even if the economic problems generated by the publication crisis are not directly solved by these actions. To reach this goal via legal means, the explicit wording of the proposal has to be carefully chosen to prevent unintended consequences and to guarantee the legal validity of the clause. In order to detect potential conflicts with existing legislation and identify any legal ambiguities, a detailed analysis of the KMK proposal's wording will be conducted in the next section. This analysis will then form the basis for legal argumentation.

[1] Referee draft of a second law governing copyright in the information society (Referentenentwurf eines zweiten Gesetzes zur Regelung des Urheberrechts in der Informationsgesellschaft).

3 Content Analysis and Legal Consequences

To elaborate upon both the meaning and the possible legal consequences of the proposal, a detailed analysis of its content and wording (where necessary) will be presented in this section.

3.1 Scope of Application

Personal scope of application

a) Employees

Affected are the employees of universities and research establishments. The condition for the application of § 43 clause 2 UrhG/KMK-P[2] is therefore the existence of an employee relationship. Employees as well as public officials (Beamte) would thereby fall inside the scope of application. § 42 Hochschulrahmengesetz (HRG) defines university personnel as primarily consisting of professors, scientific employees and teachers for special purposes. These persons would be affected by the KMK proposal in any case. Regarding the currently valid version of § 43 UrhG, the prevailing opinion is that this norm can be generally applied to all scientific university personnel (Dreier and Schulze, 2006, § 43 recital 12). Obviously, employees need not be permanent. Employees of the universities and research establishments also include student workers and, according to § 25 clause 5 HRG, most of the time scientific assistants whose salaries are usually paid by third party funding. Students who are not employed at least as assistant student workers are not affected by § 43 clause 2 UrhG/KMK-P. External PhD-students and scholarship holders do not fall within the personal scope of application because they are not employees.

b) Universities and research establishments

Based on the presentation of the KMK proposal in the KMK's statement on the draft of a second law governing copyright in the information society (Kultusministerkonferenz, 2005, p. 101 et sqq), it could be reasoned that by naming "universities and research establishments" only the publicly funded ones are meant. Privately funded universities and research establishments shall obviously not be affected. But the wording of § 43 clause 2 UrhG/KMK-P does not permit this conclusion. Therefore, adding "publicly funded" to the clause would make this clear, if this indeed was the intent.

Objective scope of application

Further affected by § 43 clause 2 UrhG/KMK-P are scientific works that were created by university personnel within the scope of their teaching and research activities.

[2] UrhG/KMK-P as a short form of UrhG de lege ferenda according to the KMK-Proposal of 2005.

a) Scientific works

First of all, in the context of German copyright law, a work is defined as a “personal mental creation” according to § 2 clause 2 UrhG (compare e.g. Dreier and Schulze 2006, § 2 Recital 13; Heermann 1999, p. 468). To qualify for protection, the work has to demonstrate a certain level of creativity as well as exist in a “recognizable form” (compare v. Moltke 1992, p. 171 et sqq). Scientific results or pure ideas are not protectable as such (Dreier and Schulze 2006, § 2 recitals 37-41; Heermann 1999, p. 468; Möhring and Nicolini - Ahlberg, 2000, § 2 recital 46). Neither are inventions; they instead fall within the scope of application of § 42 ArbNErfG (Employee Invention Law).

Beyond that, § 43 clause 2 UrhG/KMK-P aims only at the scientific works produced by university personnel. A work is considered scientific if it is the result of “a mental activity whose goal is the attainment of new knowledge in a methodological, systematic and revisable way” (Bundesverfassungsgericht, 1973, p. 113). So not only journal articles fall within the scope of application of § 43 clause 2 UrhG/KMK-P, but generally also monographs, dissertations, habilitations and other scientific writings. In contrast to journal articles, for which authors receive little or no compensation, these kinds of works may command larger sums (depending on the scientific discipline). Therefore, where these kinds of works are concerned, the authors themselves may be directly affected. With respect to journal articles, however, in most cases it is solely the interests of the publishers and universities that may come into conflict. Meanwhile, the author himself would probably be glad to see his journal article accessible via a university archive, as his results would thereby be more widely disseminated (if a publisher indeed still accepts articles under these conditions).[3] This argues for a limitation of the scope of application of any such norm to journal articles and similar publications, at least if a legal obligation is established. Hansen included such a limitation in his proposal for a new sentence 3 in § 38 clause 1 UrhG, which states that the exclusivity of the publisher’s usage right shall expire after six months, given that this is justified for non-commercial purposes (Hansen, 2005, p. 387). Harnad (2006b), as a strong proponent of open access policies, summarized some principles for university OA-mandates. This generic rationale and model also refers only to journal articles, not monographs.[4] Although one could interpret the KMK proposal in this sense, a clarification would be of great benefit here.

b) Within the scope of teaching and research activities

In addition, § 43 clause 2 UrhG/KMK-P is only to be applied if the scientific works have been produced within the scope of university employees’ teaching and research activities. As mentioned above, theses (both pre- and post-doctoral) and dissertations would be included as well. However, these would have had to be created within the

[3] According to the listing on `http://romeo.eprints.org/stats.php`, in the beginning of 2006 about 68% of the approximately 8000 journals surveyed allowed their authors make the postprint of their work publicly available through self- or institutional archiving.

[4] Also see Harnad’s comment on this in the American Scientist Open Access Forum (Harnad, 2006a).

confines of teaching and research activities. However, these terms make differentiation somewhat difficult. In a broad interpretation, virtually every scientific work could be classified as having been created within the scope of a scientist's teaching and research activities, because scientific works normally result from research activities. In a narrower interpretation of the scope of teaching and research activities, it could be interpreted as pertaining solely to the sphere of employment. In the case of employed PhD students, for example, this differentiation will probably not be simple. Even in the instance of a strictly narrow interpretation of the term "teaching and research activities", meaning its equation with "in the scope of their employment", it would have to be clarified which of the professors' activities would be covered. In general, it is assumed that the publication of research results does not belong to the official duties of the professors (In that sense e.g. (Dreier and Schulze, 2006, § 43 recital 12); (Leuze, 2003, p. 122)) and it is assumed that their works are in principle no obligatory works (or works-for-hire) but so called free works (compare (v. Moltke, 1992, p. 227); (Schricker, 1999, § 43 recital 31); (Heermann, 1999, p. 475)). Adhering to the term "teaching and research activities" in a literal sense, the works of the professors would be included in the regulation content, regardless of the context in which they were created. "In the scope of their teaching and research activities" could possibly also affect works from third party funding. In any case, there is need for clarification concerning the intention of the KMK proposal.

c) Intent to publish

If, according to § 43 clause 2 sentence 2 KMK-P, the intent to publish is to be announced without undue delay, and, irrespective of the defective wording[5], a digital copy of the work has to be provided for the university, it is crucial to determine at which point in time an intent to publish can be said to exist. In theory, a mere idea in a researcher's mind, the desire to communicate an idea in written form, or the act of penning the first sentences of an article could be interpreted as constituting an intent to publish. It can rather be assumed that the concrete intent to publish only exists after the completion of the work itself, however, at any point prior to its completion, the work cannot be said to exist at all in a form equivalent to the finished product. Particularly in scientific communication, preliminary versions typically undergo repeated revisions and often differ significantly from the final submitted version.

This view would be consistent with § 2 UrhG, which states that a work protected by copyright in form of a personal mental creation does not exist until it takes shape of any kind. In other words, it has to be perceptible (Dreier and Schulze, 2006, § 2 recital 13). Hence, the intent to publish cannot be said to be present until the work itself exists, even if the work was created with the aim of later publication. In principle, however, it is possible to grant usage rights, e.g. publication rights, prior to the execution of a work. In this context, it is only feasible to apply the portion of

[5] "and provide these in digital form" would thus equate to the intent to publish. This is certainly not what is meant here. A more correct alternative, would, for instance, be: "and to provide the relevant work in digital form."

the statute requiring the announcement of intent to publish after the work has been completed. Otherwise, both universities and authors would be confronted with far too many uncertainties. Limiting the scope of the proposed legislation to only those works intended for publication would allow authors to freely decide the fate of each article they write. The implementation of this simple limitation in this context helps to prevent the advent of complex future classification criteria for academic works (e.g. of teaching materials).

3.2 Legal Consequences

Announce intent to publish without undue delay and provide a digital copy

If the requirement for the immediate announcement of intent to publish from § 43 clause 2 sentence 2 UrhG/KMK-P were to be linked to the obligation to provide the work in digital format, this would imply that the university would have to receive a digital copy of the work before it could be submitted to a publisher. This condition would ensure that an author could not (legally) circumvent his obligation to the university by granting an exclusive right of use to a publisher at an earlier point of time before granting the obligatory right to the university. In this case, the exclusive right of use would have the consequence that the author could not grant any other usage rights to the work after that. Similar to § 9 VerlG (German publisher law), the delivery of the piece of work would be included in the obligations within the scope of granting the right of use. This obligation would there not present anything new to German copyright law.

Entitlement to a non-exclusive right of use

a) Acquisition of the right of use

According to § 43 clause 2 sentence 1 UrhG/KMK-P, universities and research establishments "are entitled to" a non-exclusive right of use. However, the legal consequences have to be specified here. "To be entitled to" could mean that the university has right to lay a claim to the usage right or that this usage right is automatically extended to it whenever a work is submitted. Another possibility would be a tacit granting of the right of use. This characterization is relevant concerning the so-called "Sukzessionsschutz" in German copyright law from § 33 sentence 1 UrhG. According to this measure, a simple right of use remains valid for the university even if usage rights are subsequently granted by the author to other institutions for the same work (compare (Dreier and Schulze, 2006, § 33 recital 1 et sqq.)) In other words, the author could grant a publisher the exclusive right of use for his work even if the university had already received a non-exclusive right of use from § 43 cl. 2 s. 1 UrhG/KMK-P.[6] In

[6] On possible liability due to insufficient information about the pre-existing simple right of use, see Dreier and Schulze (2006, § 33 recital 6) and Schricker (1999, § 33 recital 6).

this instance, the university would continue to retain its non-exclusive usage right per § 33 sentence 1 UrhG. It is the order of events that is critical here: Were the author to grant exclusive right of use to a publisher first, he would no longer be permitted to extend a non-exclusive right of use to the university. This is opposed to the notion of an exclusive right of use being a right in rem (Dreier and Schulze (2006, § 33 recital 4); also compare Schricker (1999, § 33 recital 9)). A tacit assignment of the right of use can be assumed at the latest when the author provides a digital copy of his work to the university. This actions serves to fulfill the obligation set down in § 43 cl. 2 s. 2 UhrG/KMK-P which is inclusive of the assignment of the right of use. (Compare on this Möhring and Nicolini - Spautz, 2000, § 43 recital 8; Dreier and Schulze 2006, § 43 recital 19; Schricker - Rojahn, 1999, § 43 recital 41).

b) Content of the right of use

In practice, the concrete content of the non-exclusive right of use could affect the author's ability to publish his works with commercial publishers. Therefore, this content should be more specific at this juncture. For example, it might well make a difference to a publisher interested in a given work if the author's university simply has the right to provide the work in digital form on a server or if it may additionally distribute the document via its own journals, mailing lists or anthologies. A limitation of the right of use to certain types of use and, where applicable, the definition of conditions (such as the limitation on non-commercial utilization), could help to avert possible conflicts arising from the vague wording. For example, certain publishers may demand the exclusive right of use as a condition for publishing an author's work. This type of arrangement is obviously severely limiting for the author; he would be forced to find publishers that do not impose such restrictive conditions. However, having work published in a prestigious journal is of paramount importance to a researcher; tenure and promotion processes in academia are in part predicated upon one's publication record. The Wissenschaftsrat (2005) confirms the importance that the number of publications in renowned journals has for an academic who is up for a research performance evaluation.

Thus, an explicit limitation of the university's right of use would facilitate possible negotiations on publication contracts for the authors. According to § 31 cl. 5 s. 1 UrhG, the types of use comprised in a right of use are determined by the purpose of the contract forming the basis of the right of use assignment. Therefore, if the breadth of the granted right of use is unclear, it is to be established through interpretation (Dreier and Schulze, 2006, § 31 recital 22). In this case, in § 43 cl. 2 UrhG/KMK-P, the content of the right of use to be granted is not explicitly defined; however, at issue here is not the granting contract per se but merely the legal obligation to grant the right of use.

Still, one could possibly consider an analogue application of § 31 cl. 5 UrhG, which would imply that the breadth of the relevant right of use would conform to the intent of § 43 cl. 2 UrhG/KMK-P. Even if one refuses such an analogue application, the principle of the so-called “Zweckübertragungslehre”[7] in German copyright law persists. This principle is also applicable, according to the prevailing opinion, to works falling within the scope of § 43 UrhG (Schricker - Rojahn, 1999, § 43 recital 48; Möhring and Nicolini - Spautz, 2000, § 43 recital 7; Dreier and Schulze 2006, § 43 recital 17). The “Zweckübertragungslehre” stipulates that in cases of doubt, the author only grants the rights of use to the extent necessary to fulfill the purpose of the underlying contract (Compare on this e.g. (Schricker, 1999, §§ 31/32, recital 31 et sqq.)). The KMK mentions that by virtue of its proposed norm, a non-commercial secondary publication right should be made possible ((Kultusministerkonferenz, 2005, p. 103)). This would thus also form the basis for the content of the relevant right of use. However, greater legal certainty would be afforded by an explicit definition of the usage right’s content directly in the wording of the proposal itself. This definition could be relevant vis a vis a possible conflict with the academic freedom guaranteed by article 5 clause 3 Grundgesetz (German Constitution) as well. The definition would serve to establish the breadth of a possible intervention with respect to the constitutional rights of the researcher as an author.

4 Constitutional Legitimacy

One of the criticisms concerning the KMK proposal (also discussed for example by Hansen (2005)) addresses the issue of constitutional legitimacy. Article 5 clause 3 Grundgesetz guarantees the so-called right to academic freedom. Should a legal change be made now, it has to be in line with this constitutionally guaranteed right. With respect to the KMK proposal, the freedom of publication that is derived from academic freedom (Leuze 2003, p. 122; also compare Steinfort 1987, p. 20 et sqq.) might be curtailed and thereby constitute an infringement of constitutional law. [8] The works addressed by the proposal are those created by university employees, including professors, who are holders of this basic right. In principle, this establishes an instance of applicability. Generally, the scientist has the right to decide where, when and how to publish his results (Fehling, 2005, art. 5 clause III recital 74). This is referred to as “positive” publication freedom. In addition, the academic freedom comprises the so called negative publication freedom: the right to decide if one’s results should be published at all. These rights can be derived from article 5 clause 3 of the German constitution which guarantees the so-called academic freedom and influences all legal rules touching the holders of this right within the scope of application. Another norm that guarantees these rights is part of the author’s moral rights. § 12 clause 1 UrhG would also be in conflict with the KMK proposal (see also Hansen 2005, p. 379), but this conflict might be solved by introducing § 43 clause 2 UrhG/KMK-P as a lex specialis of § 12 UrhG.

[7] Compare on this e.g. Schricker (1999, §§ 31/32, recital 31 et sqq.)

[8] On the constitutional legitimacy of a first publication right for the university, see Pflüger and Ertmann (2004, p. 441); and from a critical perspective, Hansen (2005, p. 387).

Should authors be legally required to grant their universities a non-exclusive right of use for their work, their right to positive publication freedom could be compromised (assuming that the authors do indeed intend to publish their work). In this case, the author can generally still decide where to submit his work for publication. Yet in practice, he can legally only publish the work with publishers who do not insist on an exclusive right of use. The author would then have to find another publisher who is willing to publish his work regardless of the university's usage rights.

If the author grants the publisher an exclusive right of use before the university is granted a non-exclusive one, then the author can no longer fulfill his legal obligation from § 43 Abs. 2 UrhG/KMK-P. But this would probably only be possible if the author had already neglected his duty to provide a digital copy of his work without undue delay. This problem could be resolved in the interest of the university by stipulating that the copy and the right of use must be granted to it before the work is submitted to any publisher. In turn, the university would be permitted to invoke this right only when the work is actually published and not before. similar stipulation can be found in the proposal for the U.S. CURES Act (2005) whereby a repository deposit of publicly funded scientific works at the time of journal acceptance is required.

If only the German market for scientific information is considered, it could be argued that because every publication by German professors and researchers at public institutions would be subject to these conditions, the publishers should have to change their publication conditions. However, as long as this market is considered to be international, German researchers might be at some disadvantage compared to their colleagues abroad when submitting their works to internationally active publishers. This does in fact constitute a limitation of the author's freedom to decide where to publish his works, or, in other words, an infringement of his right to positive publication freedom. In contrast, the decision of if and when to publish the work would remain in the hands of the academic, thereby preserving his right to negative publication freedom; only works that are already stated for publication fall within the scope of this norm.

Meanwhile, Pfüger and Ertmann recommend requiring university employees to offer first (non-exclusive) rights to any and all scientific works created within the scope of their teaching and research activities to the university. Because it requires offering the work to the universities first, this arrangement would certainly curtail the author's positive publication freedom to decide where and when to submit his work. Furthermore, in this case negative publication freedom would likewise be impacted since there is no such limitation on works intended to be published with certainty (See also Hansen 2005, p. 379). An intervention in academic or publication freedoms may be valid if it is deemed necessary and proper for attaining a goal that is generally derivable from the constitution (Classen, 1994, p. 124). The grade of this interference with respect to the decision of where and how to publish is in practice probably dependent on the overall negotiating positions of publishers and authors in the market.

Some argue that the publishers will change their conditions as soon as the legal mandate for a university right of use becomes valid. This forecast, along with other economic consequences, cannot be predicted with certainty. In disciplines which are very much concentrated on national communication, such as jurisprudence, publishers might be more willing to cooperate than in other disciplines with stronger international communication structures. In these disciplines, university members might have to contend with more disadvantages compared to their foreign colleagues if the lack of an international approach persists. Justification of the intervention will therefore be difficult, whereas the goal of providing increased access to scientific literature as the basis for new research can, in our opinion, be derived from constitutional academic freedom itself. The holder of this right has a claim on both the framework and possibilities necessary to conduct research as an expression of his academic freedom.

5 Conclusion

At the moment, there seems to be no simple solution for the problem of introducing a shift in scientific publishing in order to promote wider access to scientific knowledge. If new legislation is introduced, there are several marginal conditions as well as practical consequences that have to be considered. The KMK proposal to add a new § 43 clause 2 UrhG to German copyright law creates some problems concerning its constitutional legitimacy. This is because it impinges upon academic freedom, and it might be difficult to justify this without having a sufficient grade of probability for the forecasting of the consequences. In addition, there are vague aspects that ought to be clarified. Overall, the KMK proposal relaxes the objections to e.g. the recommendation for an offering obligation proposed by Pflüger/Ertmann, but still leaves too many open questions.

Meanwhile, Hansen's recommendations with respect to contractual copyright law are less problematic in terms of legitimacy and practicability. His proposal would entail the expiration of any exclusivity of a right of use after six months and has been adopted (with few changes) by the Bundesrat in its comment on the draft of a second law governing copyright in the information society on May 19th 2006. A straightforward approach to legal change in this arena would be to formulate an accordant statute of limitation for copyrights. This would need to be valid in light of the Three-Step-Test, a protocol that stipulates criteria for the national legislation of the European Union member states that have to be met. SLimitations on exclusive rights of use could also be helpful in combination with institutional or funding policies promoting self-archiving of publications. Generally, the problem does not seem to lie in any reluctance on the part of the researchers to make their work openly accessible. Rather, at issue appears to be a lack of opportunities to do so as well as insufficient awareness about the topic (Fournier 2005, p. 235 et sqq. with more evidence). Therefore, legal changes might be more effective if they are aimed at the provision of possibilities rather than restricting the publishing behavior of publicly funded researchers. Nonetheless, one challenge will be to make the advantages of new forms of scientific communication possible without loss of quality. Another will be to apply as little external force as possible to the market as long as the economic consequences of certain changes are still so vague.

References

Berlin Declaration (2003): "Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities," `http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf`.

Bundesverfassungsgericht (1973): "BVerfGE 35, 75ff," .

Classen, C. D. (1994): "Wissenschaftsfreiheit außerhalb der Hochschule," in: *Tübinger rechtswissenschaftliche Abhandlungen Nr. 77*, Mohr, Tübingen.

Dreier, T. and G. Schulze (2006): *Urheberrechtsgesetz Kommentar*, 2nd edition, Beck, München.

Fehling, M. (2005): in: Dolzer, Vogel, and Graßhof (eds.), *Bonner Kommentar zum Grundgesetz*, C. F. Müller, Heidelberg.

Fournier, J. (2005): "Zur Bedeutung von Open Access für das Publikationsverhalten DFG-geförderter Wissenschaftler," *Zeitschrift für Bibliothekswesen und Bibliographie*, pp. 235–244.

Hansen, G. (2005): "Zugang zu wissenschaftlicher Information - alternative urheberrechtliche Ansätze," *Gewerblicher Rechtsschutz und Urheberrecht Internationaler Teil*, pp. 378–388.

Harnad, S. (2006a): "American Scientist Open Access Forum (AmSciOAF)," available at http://www.ecs.soton.ac.uk/ harnad/Hypermail/Amsci/5214.html, 12.03.2006.

Harnad, S. (2006b): "Generic Rationale and Model for University Open Access Self-Archiving Mandate," available at http://eprints.ecs.soton.ac.uk/12078/, 13.03.2006.

Heermann, P. W. (1999): "Der Schutzumfang von Sprachwerken der Wissenschaft und die urheberrechtliche Stellung von Hochschulangehörigen," *Gewerblicher Rechtsschutz und Urheberrecht*, pp. 468–476.

Kultusministerkonferenz (2005): "Stellungnahme der Kultusministerkonferenz zum Referentenentwurf eines zweiten Gesetzes zur Regelung des Urheberrechts in der Informationsgesellschaft in der Fassung vom 27.9.2004," in: Hoeren and Sieber (eds.), *Urheberrecht für Bildung und Wissenschaft*, vol. 2, Hochschulrektorenkonferenz (Beiträge zur Hochschulpolitik), p. 88 et sqq.

Leuze, D. (2003): *Urheberrechte der Beschäftigten im öffentlichen Dienst*, 2nd edition, Erich Schmidt Verlag, Berlin.

Mittler, E. (2005): "Gefahren von vorgesehenen Bestimmungen des 2. Korbs der Ur-heberrechtsänderungen für die wissenschaftliche Literaturversorgung," in: Hoeren and Sieber (eds.), *Urheberrecht für Bildung und Wissenschaft, Hochschulrektorenkonferenz (Beiträge zur Hochschulpolitik, vol. 2)*, p. 44 et sqq.

Möhring, P. and K. Nicolini (eds.) (2000): *Urheberrechtsgesetz Kommentar*, 2nd edition, Vahlen, München.

Pflüger, T. and D. Ertmann (2004): "E-Publishing und Open Access – Konsequenzen für das Urheberrecht im Hochschulbereich," *ZUM*, pp. 436–443.

Schricker, G. (ed.) (1999): *Urheberrecht Kommentar*, 2nd edition, Beck, München.

Steinfort, F. (1987): *Die verfassungsrechtlichen Grundlagen der Veröffentlichungsfreiheit des Wissenschaftlers*, Dissertation, Bonn.

U.S. CURES Act (2005): “Bill for the American Center for Cures Act,” introduced 12/14/2005, 109th congress, 1st session, S.2104, Section 499H-1.

v. Moltke, B. (1992): *Das Urheberrecht an den Werken der Wissenschaft*, Nomos, Baden-Baden.

Wissenschaftsrat (2005): “Empfehlungen zur Ausgestaltung von Berufungsverfahren,” Drucksache 6709-05.

Part II

Business Structure

Infrastructure

Microstructure

Yield Management: Beyond Expected Revenue

Christiane Barz[1], Marliese Uhrig-Homburg[2], and Karl-Heinz Waldmann[1]

[1] Institute for Economic Theory and Operations Research,
Universität Karlsruhe (TH)
{barz,waldmann}@wior.uni-karlsruhe.de

[2] Institute for Finance, Banking and Insurance,
Universität Karlsruhe (TH)
uhrig@fbv.uni-karlsruhe.de

Summary. Yield management models are usually solved for an optimal, i.e. expected revenue maximizing, policy. However, the consequences of a certain yield management practice affect not only the expected revenue of a flight, but also factors influencing future revenues. In certain business scenarios, these consequences can be crucial and should be considered in the capacity allocation process. Therefore, this paper promotes the use of a more sophisticated optimality criterion for yield management problems. As a first step, some simple optimality criteria are suggested and compared in an example.

1 Introduction

Yield management, also known as revenue management, is the process of understanding, anticipating, and reacting to consumer behavior in order to maximize expected revenue. According to Cross (1998), it ensures that companies allocate the right type of capacity to the right kind of customer at the right time for the right price.

The most prominent example of an industry using yield management is the airline industry. Airlines typically divide a pool of identical seats into several booking classes that represent e.g. different discount levels with differentiated sale conditions and restrictions. Mixing discount and higher-fare passengers in the same aircraft compartment offers the airline the potential of gaining revenue from seats that would otherwise fly empty. The optimal allocation of capacity to different classes of demand forms a substantial part of a yield management system.

In a setting with high fixed costs and negligible variable costs, as in the airline industry, it is often argued that maximizing expected revenue of a single flight is an appropriate objective for finding the optimal mix of high-fare and discount demand. However, the consequences of a certain yield management practice affect not only the expected revenue of this flight, but also factors influencing future revenues. Expected passenger spill, i.e. the number of full-fare reservation requests that must be turned away, or the expected number of denied boardings, might cause customer dissatisfaction and reduce future demand. The expected load factor as a metric for market share as well as the variability of revenues might be of interest to investors. In certain business scenarios, these consequences can be crucial and should be considered in the capacity allocation process.

In the present paper, we turn our attention to these frequently overlooked factors in yield management. For convenience, we will stick to the terminology of the airline industry and restrict attention to the static model without cancellations, no-shows or group requests, since it is the simplest and most widespread yield management model. However, the ideas to be outlined are independent of the (expected revenue maximizing) yield management model used and all ideas can of course be applied to other industries using yield management accordingly.

The paper is organized as follows. In Section 2, we give a brief review of the static yield management model. Then, in Section 3, we introduce a more general optimization criterion for yield management and provide an overview of the related literature. For special cases, optimality equations for finding a more appropriate yield management policy and structures of this policy are given in Section 4.

2 The Static Yield Management Model

The basic setting of all single-leg yield management models considers a nonstop flight of an airplane with a capacity of C seats that is to depart after a certain time horizon. There are k ($k \in \mathbb{N}$) booking classes with positive fares ordered according to $\hat{r}_1 \geq \hat{r}_2 \geq \cdots \geq \hat{r}_k$.

In the early literature on seat inventory control, the demand of each booking class is supposed to arrive during a single contiguous time segment. In this case, the booking period is divided into periods with booking requests belonging to the same fare class. As soon as the total demand of a booking class is known, the amount of demand that has to be accepted in order to maximize the expected revenue of that flight must be determined. The total demands of the booking classes are assumed to be independent random variables $D_1, \ldots, D_k$. Neither cancellations nor no-shows are allowed.

Most of these static models additionally assume that customer requests for tickets arrive in increasing fare order, i.e. the class willing to pay the fare $\hat{r}_k$ before $\hat{r}_{k-1}$, etc. For simplicity, we will stick to this assumption in the following.

The static yield management model with two fare classes was first introduced by Littlewood (1972) and extended heuristically to more than two fare classes by Belobaba (1987) and Belobaba and Weatherford (1996). An exact solution can be found in Curry (1990), Wollmer (1992), Brumelle and McGill (1993) and Robinson (1995).

The objective of finding a policy maximizing the expected revenue can be reduced to solving the optimality equation of a finite stage Markov decision model $MDP(k, \mathfrak{X}, \mathfrak{A}, (q_i), (r_i), V_0)$ with planning horizon k, state space $\mathfrak{X} = \{(c, d) \in \mathbb{Z} \times \mathbb{N}_0 \mid c \leq C\}$, where we refer to c as the remaining capacity and to d as the demand observed for the actual booking class, set $A(c, d) = \{0, \ldots, d\}$ of admissible actions in state (c, d), transition law q_i for $i = k, k-1, \ldots, 1$ such that $q_i((c, d), a, (c - a, d')) =$

$P(D_{i-1} = d')$ and 0 otherwise, one-stage rewards $r_i((c,d),a) = a \cdot \hat{r}_i$, and terminal reward $V_0((c,d)) = 0$ for $c \geq 0$ and $V_0((c,d)) = \bar{r} \cdot c$ for $c < 0$ with $\bar{r} > \max_i\{\hat{r}_i\}$. Thus, for $i = k, k-1, \ldots, 1$ and all booking classes i, given the residual capacity c_i and demand d_i, we have to determine the number $a_i = f_i(c_i, d_i) \in \{0, \ldots, d_i\}$ of seats to be accepted.

A (Markov) policy $\pi = (f_k, f_{k-1}, \ldots, f_1)$ is then defined as a sequence $f_k, f_{k-1}, \ldots, f_1$ of decision rules f_i specifying the action $a_i = f_i(c_i, d_i)$ to be taken at stage i in state (c_i, d_i). Let F denote the set of all decision rules and F^k the set of all policies.

Denote by $(X_k, X_{k-1}, \ldots, X_0)$ the state process of the MDP and introduce

$$V^*(c,d) = \max_{\pi \in F^k} E_\pi \left[\sum_{i=1}^{k} r_i(X_i, f_i(X_i)) + V_0(X_0) \mid X_k = (c,d) \right], \quad (c,d) \in \mathfrak{X}, \qquad (1)$$

as the maximal expected revenue; the random variable $\sum_{i=1}^{k} r_i(X_i, f_i(X_i))$ will be called revenue in the sequel. It is well known in dynamic programming that $V^* \equiv V_k$ is the unique solution to the optimality equation

$$V_i(c,d) = \max_{a=0,\ldots,d} \left\{ a \cdot \hat{r}_i + \sum_{d'=0}^{\infty} P(D_{i-1} = d') V_{i-1}(c-a, d') \right\}, \qquad (2)$$

which can be obtained for $i = 1, \ldots, k$ by backward induction starting with V_0. Moreover, each policy π^* formed by actions $a^* = f_i^*(c,d)$ each maximizing the right hand side of (2) is optimal, i.e. leads to V^*.

It is widely known (see e.g. Wollmer (1992), Lautenbacher and Stidham (1999), Talluri and van Ryzin (2004, p. 36-40)) that $f_i^*(\cdot, d)$ is monotone in the remaining capacity c, i.e. the more capacity is available, the more one is willing to sell. Further, for fixed capacity c and observed demand d, the optimal policy π^* is monotone in the remaining arrival periods i. Finally, it can be shown that f_i^* can be described in terms of (so-called) protection levels y_i^* such that $f_i^*(c,d) = \min\{(c - y_{i-1}^*)^+, d\}$, where $z^+ := \max\{z, 0\}$ for $z \in \mathbb{R}$. Thus, in each arrival period i a number of y_{i-1}^* seats has to be reserved for demand from classes $i-1, \ldots, 1$.

For an illustration, consider the following data taken from Belobaba (1987): There are 4 fare classes with fare prices of $\hat{r}_1 = 105 \geq \hat{r}_2 = 83 \geq \hat{r}_3 = 57 \geq \hat{r}_4 = 39$. The total capacity is $C = 107$. The demand is normally distributed (rounded to integer values). Table 1 shows the associated expectations and standard deviations.

The protection levels obtained by solving (2) read: $y_3^* = 77$, $y_2^* = 49$, and $y_1^* = 13$. This means that e.g. 77 seats are protected for classes 1, 2 and 3 and at most $107 - 77 = 30$ seats would be sold to class 4 customers.

Note that by applying π^*, the expected revenue is 6537.3 and the standard deviation of revenue earned is 1149.2. If we let c_0 denote the residual capacity at time 0, the expected load factor is $E_{\pi^*}[(C - c_0)/C] = 0.8695$, and the expected spill rate of class 1 customers, $s' = (d_1 - c_1)^+/d_1$ for $d_1 \neq 0$, and $s' = 0$ for $d_1 = 0$, is $E_{\pi^*}[s'] = 0.0941$.

Table 1: Parameters of the normally distributed demands

fare class i	$E[D_i]$	$\sigma[D_i]$
1	20.3	8.6
2	33.4	15.1
3	19.3	9.2
4	29.7	13.1

Details on the interpretation of the expected load factor and expected spill rates will be given in Section 4.

3 Expected Utility Maximization

It is obvious that the decision whether to accept or deny a customer request has an impact not only on the expected revenue but also on the whole distribution of revenue earned on that flight as well as on the load factor, the spill rates, and the passenger mix. In case of cancellations and overbooking, denied boardings are another attribute of the decision actually taken.

We assume that the main business objective is maximizing expected (long-term) profit. Since variable costs are assumed to be negligible, it is often concluded that maximizing expected revenue is an appropriate equivalent objective. While this is true in a non-changing environment (of demand and fixed costs), this equivalence is to be questioned if we assume that the accept/deny-decisions maximizing revenue affect these environmental factors. Lost customer goodwill, e.g. due to high spill rates, may lower demand for future flights (see e.g. Lindenmeier and Tscheulin (2005), Wirtz et al. (2003)). Disappointed investors, e.g. due to a reduction of market share (determined by load factors) or volatile revenues earned, might increase the costs of outside capital in the long-run or even cause venture capitalists to leave the company.

Thus, there may be other important business objectives in addition to maximizing expected revenue. Choosing a yield management policy by a sole comparison of expected revenue might be inadequate for meeting them.

3.1 A Utility Framework

We assume that the decision-maker can specify his random variables (usually referred to as attributes) of interest $\{O_1, \dots, O_m\}$ and value each possible realization of these attributes with a von Neumann-Morgenstern utility function $u(o_1, \dots, o_m)$. The policy that is preferred by the decision-maker is the policy π^* with

$$E_{\pi^*}[u(O_1, \dots O_m)] = \max_{\pi \in \Pi} E_\pi[u(O_1, \dots O_m)]$$

and Π being a set of admissible policies.

Remember that the attributes $O_1, \ldots O_m$ could stand for the revenue earned, the load factor, spill rates, etc. and u is a von-Neumann Morgenstern utility function defined on possible realizations of these random variables. Thus, $E_\pi[u(O_1, \ldots O_m)]$ reveals the true preferences of the decision-maker over different yield management policies. Expected revenue might determine (part of) these preferences, but needs not to be the only relevant feature. Thus, approximating the true preferences by expected revenue might not result in the most preferred policy with respect to the (real) utility function.

In the following, our aim will be to find yield management policies that consider the true preferences maximizing expected utility instead of expected revenue.

3.2 Related Literature

To the authors' knowledge, an overall assessment of expected utility maximization (beyond expected revenue maximization) in yield management has not been conducted up to now. Certain parts of the problem, however, have been discussed to a varying degree.

Considering higher moments of the revenue earned is a frequently mentioned, but not yet thoroughly discussed extension of existing yield management models.

Lancaster (2003) emphasized the need of risk-aware yield management within the airline industry from a controlling perspective, underlining the impact of revenue managers on the revenue distribution. Weatherford (2004) proposed an ad-hoc extension of Belobaba's EMSR-b heuristic for finding policies with a lower variation of revenue. Barz (2006) formulated a corresponding problem that penalizes this variation by using a concave utility function on revenue and provided a solution for the static k-class model; structural results were given in Barz and Waldmann (2006). For the special case of network traffic control, Mitra and Wang (2005) considered a mean risk model with the revenue's variance and tail revenue at risk as a risk-measure.

The importance of high-value customer goodwill is a more mature secondary (yield) management objective (see e.g. Weatherford and Bodily (1992), Lindenmeier and Tscheulin (2005) and Wirtz et al. (2003)). The importance of high-class passenger spill, i.e. the number of full-fare reservation requests that must be turned away, was mentioned even in the very first paper on yield management, in Littlewood (1972). Brumelle et al. (1990) included this objective by introducing a goodwill premium that is added to the highest fare class in their static two class model. This premium increases the full-fare protection level and eventually reduces the full-fare spill rate.

Similarly, combined yield management and overbooking models (e.g. Subramanian et al. (1999)) value a denied boarding by a revenue loss that might be higher than the monetary costs of the immediate consequences. Although these models' optimality criterion is formulated in terms of maximizing revenue only, in this way they implicitly consider the possible loss of customer goodwill due to denied boardings.

4 Special Utility Functions

In order to investigate some possible consequences of maximizing $E_\pi[u(O_1, \dots O_m)]$ and to be able to formulate an adequate optimality equation, we consider the maximization of three utility functions that account for more than expected revenue. These utility functions aim at (1) maximizing the load factor and revenue at the same time, (2) minimizing passenger spill and maximizing revenue at the same time, and (3) maximizing the expected utility of revenue given an exponential utility function.

As a benchmark for comparisons, we will use a utility function that is a linear transformation of the revenue earned on a flight r,

$$u_0(r) = \alpha \cdot r + \beta \quad \text{with } \alpha > 0,\ \beta \in \mathbb{R}.$$

In this case, maximizing expected utility corresponds to maximizing expected revenue. The objective of finding an expected revenue maximizing policy was discussed in Section 2.

4.1 Maximizing the Load Factor

In the case of a start-up, the goal of maximizing market share can be as important as maximizing (expected) revenue. This is because the greater the company's market share, the greater the awareness and acceptance of their products, which translates into lower marketing and sales expenses per dollar of revenue. And because of the widespread use of market share as an indicator of future profitability, this metric is of high interest to investors and might be crucial for future financing.

Let us assume that the company considered has decided to control the load factor (and thus market share) in addition to revenue earned. Preferences are strictly increasing in revenue and load factor. For simplicity, we assume that the attributes load factor and revenue are additive utility independent to the company in the sense of Keeney and Raiffa (1976), i.e. the utility gain by increasing revenue from a given level r to r' given the same load factor l is independent of l, the utility gain by increasing the load factor from a given level l to l' given the same revenue r is independent of r, and the company is indifferent between a 50-50 chance of obtaining either (r, l) or (r', l') and a 50-50 chance of obtaining either (r, l') or (r', l). So we assume

$$u(r, l) = u_R(r) + u_L(l). \tag{3}$$

Given the total capacity of the plane C, the load factor is a function of the number of seats unsold at departure c_0, namely $(C - c_0)/C$. Therefore, the utility function $u_L(l)$ can also be stated in terms of capacity at time 0, c_0. Since c is part of the state of the underlying MDP, this utility function $u_L((C - c_0)/C)$ can be seen as a terminal reward function. For ease of notation we introduce a utility function on c with $u_L^C(c) := u_L((C - c)/C)$.

If $u_R(r) = \alpha \cdot r + \beta$ with $\alpha > 0$ and $\beta \in \mathbb{R}$ is an increasing linear function, the optimization problem again corresponds to the optimization problem (1) given in

Section 2 if one extends the terminal reward function $V_0((c,d))$ to include u_L, so that $V_0((c,d)) = \frac{1}{\alpha} \cdot u_L^C(c) = \frac{1}{\alpha} \cdot u_L((C-c)/C)$ for $c \geq 0$ and $V_0((c,d)) = c \cdot \bar{r}$ for $c < 0$. This problem can be solved as indicated in Section 2.

By generalizing the proofs of Lautenbacher and Stidham (1999) and Talluri and van Ryzin (2004, p. 36-40), it is clear that the structural properties remain true if $u_L^C(c)$ is concave and decreasing in c and $u_L^C(0) = 0$. In this case, the optimal policy will be of protection level type and the protection levels will be increasing in the periods until departure.

If we take our example from Section 2 and try to find a policy maximizing (3) with $u_R(r) = r$ and $u_L^C(c) = -z \cdot c$ for different values of z, this corresponds to the objectives of maximizing expected revenue and expected load factor with different weights put on the importance of these objectives. In this special case, it is easy to show that protection levels decrease for increasing z.

Figure 1 shows values of the expectation of revenue, the standard deviation of revenue, the expected load factor and the expected class 1 spill rate in our example given the optimal policy for parameter $z = 0, 2, \ldots, 200$.

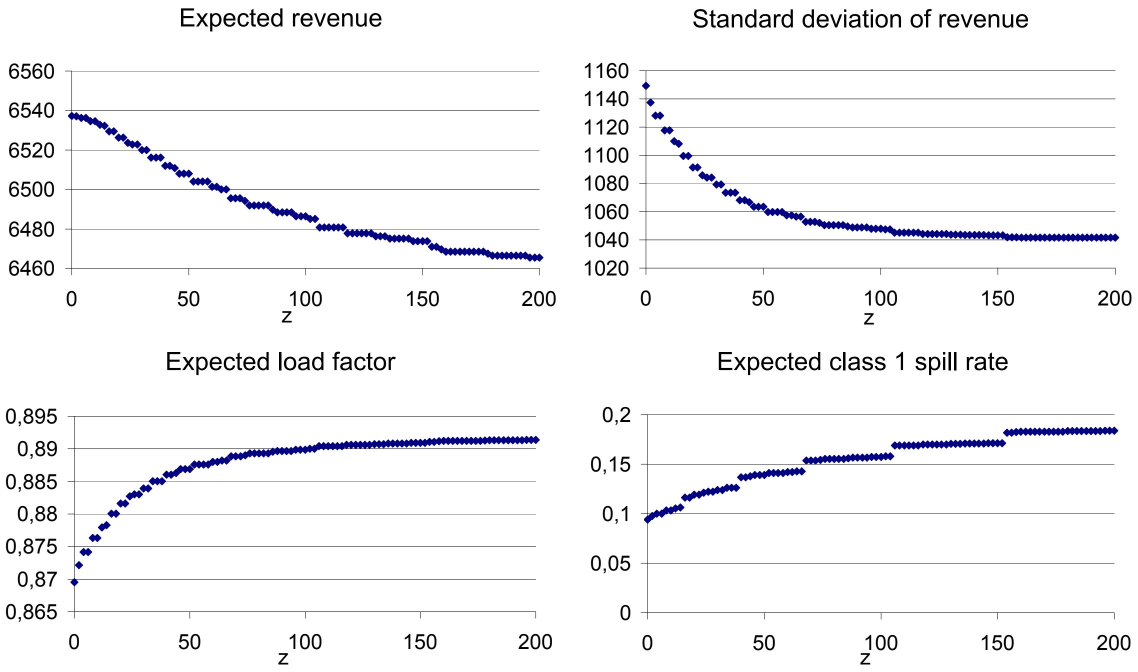


Figure 1: Plots of the expectation and standard deviation of revenue, the expected load factor, and the expected spill factor for parameter $z = 0, 2, \ldots, 200$ (each based on the optimal policy).

It is not surprising that the expected load factor increases when we put more importance on this criterion. Accordingly, expected revenue decreases with z. Since higher values of z cause lower protection levels, fewer seats will be protected for high-revenue demand and expected spill rates will increase. The standard deviation decreases because the standard deviation of the uncontrolled, first-come-first-served process is lower than that of its controlled, revenue-maximizing counterpart.

4.2 Minimizing Passenger Spill

According to Brumelle et al. (1990), ”airlines are justifiably concerned about the impact of discount seat allocation policies on the number of full fare reservation requests that must be turned away.” In the airline industry, this number (expressed as a proportion of total full fare demand) is called passenger spill (rate). Spill rates, measuring a source of decreasing customer goodwill, can be important metrics e.g. in the case of a company facing low-cost competition when the loyalty of high-value customers becomes increasingly important. Customers who are turned down for this flight might choose to make a booking on the next flight of the same company, but it is also likely that these customers will choose a competitor’s flight and will not return for future flights.

Let us therefore assume that the company under consideration has decided to control the class 1 passenger spill in addition to revenue earned. Preferences are strictly increasing in revenue and decreasing in the spill rate. For simplicity, we assume that the attributes spill and revenue are additive utility independent to the company. So we assume

$$u(r, s) = u_R(r) + u_S(s). \tag{4}$$

Given the number of unsold seats before the arrival of the last, highest-valued, customer group c_1 and the demand of this customer group d_1, the spill can be calculated by $s = (d_1 - c_1)^+$.

Thus, if $u_R(r) = \alpha \cdot r + \beta$ with $\alpha > 0$, $\beta \in \mathbb{R}$ is an increasing linear function, the optimization problem corresponds to one similar to (1) with modified one-stage reward $\tilde{r}_1((c,d),a) := r_1((c,d),a) + \frac{1}{\alpha} \cdot u_S((d_1 - c_1)^+)$, i.e. for all $(c,d) \in \mathfrak{X}$

$$W^*(c,d) = \max_{\pi \in F^k} E_\pi \left[\sum_{i=2}^{k} r_i(X_i, f_i(X_i)) + \tilde{r}_1(X_1, f_1(X_1)) + V_0(X_0) \mid X_k = (c,d) \right],$$

which can be solved as indicated in Section 2.

If, for fixed d_1, $u_S((d_1 - c_1)^+)$ is a concave function of c_1, the proofs of Talluri and van Ryzin (2004) can again be easily generalized. One is thus able to show that the optimal policy is of protection level type and that these protection levels are increasing in the number of periods until departure.

Brumelle et al. (1990) aim to reduce the spill rate by adding an additional amount z to the fare of the highest fare class r_1. In this way, the value of accepting members of this customer class is increased, and due to the revenue maximization procedure, the spill rate decreases. This ends up in the same set of strategies as in our setting with a utility function $u_S(s) = -z \cdot s$.

If we take our example from Section 2 and try to find a policy maximizing (4) with $u_R(r) = r$ and $u_S(s) = -z \cdot s$ for different values of z, this corresponds to the objectives of maximizing expected revenue and expected spill with different weights put on the importance of these objectives. In this special case, it is easy to show that protection levels increase for increasing z.

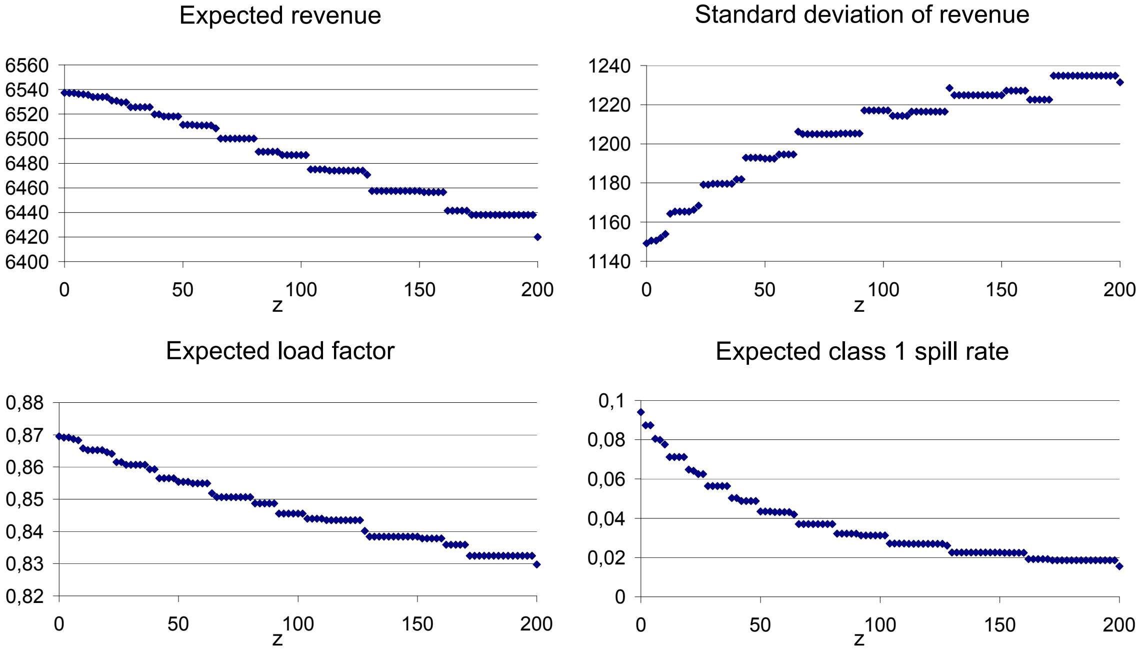


Figure 2: Plots of the expectation and standard deviation of revenue, the expected load factor, and the expected spill factor for parameter $z = 0, 2, \ldots, 200$ (each based on the optimal policy).

Based on the data of our example, Figure 2 shows expectation and standard deviation of revenue, expected load factor and expected spill factor for different values of the parameter z.

It is not surprising that the expected spill rate decreases when we put more importance on spill minimization. Accordingly, expected revenue decreases with z. Since higher values of z cause higher protection levels, more seats will be protected for high-revenue demand and so expected load factors decrease. The standard deviation tends to increase, since with higher protection levels, the revenue earned depends more and more solely on the realization of class 1 demand.

4.3 Maximizing Expected Exponential Utility of Revenue

Let us assume that the decision-maker is experiencing liquidity problems and has limited access to the capital market. Investors cast a critical eye on negative revenue variations. Customer goodwill and market share are not crucial compared to the importance of revenue and meeting investors' revenue expectations. Thus, we can simplify the goal of maximizing $u(o_1, \ldots o_m)$ to $\max u_R(r)$.

In this situation, the utility of a certain revenue earned r should be higher than that of a lottery with the expected outcome of r. Or, stated differently: For being indifferent between a revenue of r_s with probability 1 and a lottery over revenues R, the expected payoff of the lottery needs to be higher than the certain revenue, i.e.

$$E[R] = r_s + r_p, \qquad r_p > 0. \tag{5}$$

This preference structure can be modeled by a concave utility function. Kirkwood (2004) shows that in most cases an appropriately chosen exponential utility function is a very good approximation for general utility functions. This is why, as a first step, we will restrict ourselves to models with an exponential utility function with parameter $\gamma > 0$,

$$u_r^\gamma(r) = -\exp(-\gamma \cdot r).$$

The parameter γ determines the size of r_p needed for indifference in (5). The higher γ, the higher r_p is. Spoken in terms of von Neumann-Morgenstern utility functions, this function exhibits constant absolute risk aversion of γ. Thus, a decision-maker with utility function $u_r^{\gamma_1}$ of this type is more risk-averse than one with $u_r^{\gamma_2}$ in the sense of Pratt (1964) if γ_1 is bigger than γ_2.

Maximization of the expected exponential utility $U^*(c,d)$, $(c,d) \in \mathfrak{X}$,

$$U^*(c,d) = \max_{\pi \in F^k} E_\pi \left[-\exp\left(-\gamma \cdot \left[\sum_{i=1}^{k} r_i(X_i, f_i(X_i)) + V_0(X_0) \right] \right) \mid X_k = (c,d) \right],$$

is known as risk-sensitive optimality within the setting of *MDP*s and was first introduced by Howard and Matheson (1972). The resulting optimality equation can be stated as

$$U_i(c,d) = \max_{a=0,\dots,d} \left\{ \exp(-\gamma \cdot a \cdot \hat{r}_i) \cdot \sum_{d'=0}^{\infty} P(D_{i-1} = d') U_{i-1}(c-a, d') \right\}$$

with terminal reward $U_0(c,d) = -1$ for $c \geq 0$, and $U_0(c,d) = \exp(-\gamma c \bar{r})$ for $c < 0$ $(d \geq 0)$.

For $\gamma \to 0$ we again obtain the maximization of expected revenue, while $\gamma \to \infty$ yields a worst-case revenue maximization (see Coraluppi (1997)).

This type of optimality criterion has been discussed in Barz and Waldmann (2006). It turns out that the optimal policy is again of protection level type and increasing in the number of periods until departure. Simulations (cf. Barz (2006)) show that increasing values of γ result in decreasing protection levels.

The resulting measures of performance are displayed in Figure 3 for different values of γ.

It is not surprising that expected revenue decreases for increasing values of γ. Accordingly, the standard deviation decreases and converges to the level of the uncontrolled process. Since higher values of γ result in lower protection levels, fewer seats will be protected for high-revenue demand, and so expected load factors and expected spill rates will increase.

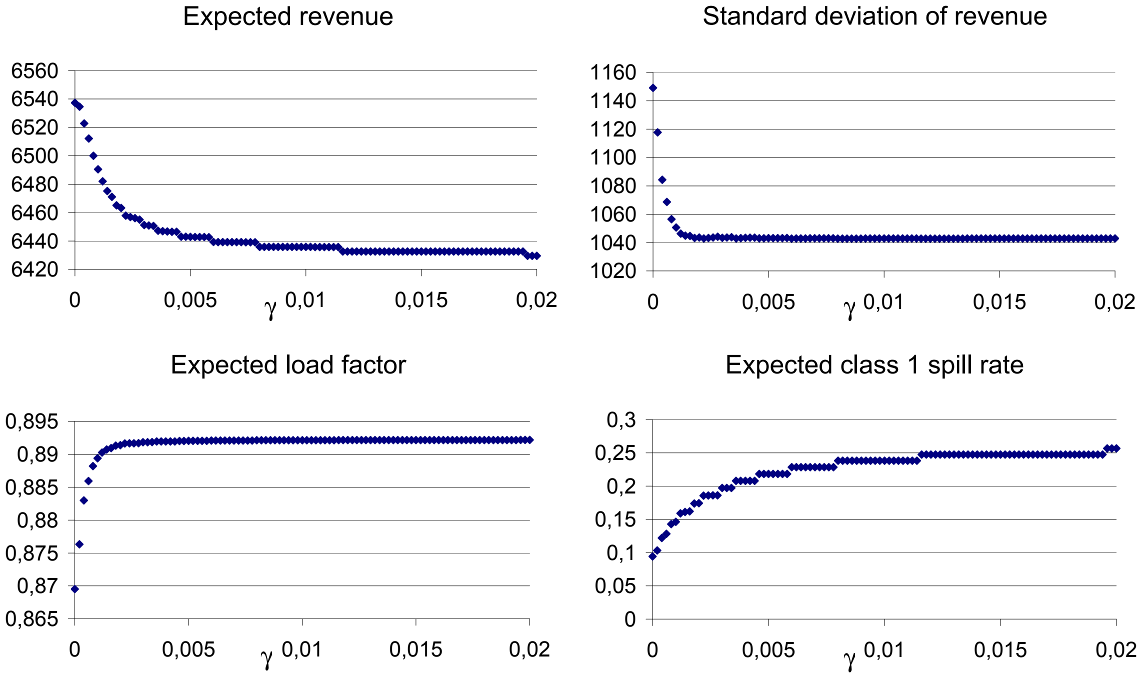


Figure 3: Plots of the expectation and standard deviation of revenue, the expected load factor, and the expected spill factor for parameter $\gamma = 0.001, 0.002, \ldots, 0.02$ (each based on the optimal policy).

5 Outlook

This paper promotes the use of a more sophisticated optimality criterion for yield management problems taking into account other factors than expected revenue. As a first step, some simple optimality criteria were suggested and compared in an example.

In addition to a thoroughly analytical investigation on possible generalizations of the suggested Bellman equations, some areas for further study may include the study of an intertemporal yield management model with more than one flight departure in order to deviate an adequate Bellman equation for the one-departure case. The incorporation of cancellations, no-shows and more sophisticated customer arrival processes into the suggested models is another possible avenue for future research.

References

Barz, C. (2006): "How does risk aversion effect optimal revenue management policies? - A simulation study," in: D. C. Mattfeld and L. Suhl (eds.), *Informationssysteme in Transport und Verkehr*, vol. 4, Books on Demand GmbH, Norderstedt, pp. 161–172.

Barz, C. and K.-H. Waldmann (2006): "Risk-sensitive capacity control in revenue management," *Submitted for publication.*

Belobaba, P. P. (1987): *Air travel demand and airline seat inventory management*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge.

Belobaba, P. P. and L. R. Weatherford (1996): “Comparing decision rules that incorporate customer diversion in perishable asset revenue management situations,” *Decision Analysis*, 27(2), pp. 343–363.

Brumelle, S. L. and J. I. McGill (1993): “Airline seat allocation with multiple nested fare classes,” *Operations Research*, 41(1), pp. 127–137.

Brumelle, S. L., J. I. McGill, T. H. Oum, K. Sawaki, and M. W. Tretheway (1990): “Allocation of airline seats between stochastically dependent demands,” *Transportation Science*, 24(3), pp. 183–192.

Coraluppi, S. P. (1997): *Optimal control of Markov decision processes for performance and robustness*, Ph.D. thesis, University of Maryland.

Cross, R. (1998): *Revenue management: hard-core tactics for profit-making and market domination*, Orion Business, London.

Curry, R. E. (1990): “Optimal airline seat allocation with fare classes nested by origins and destinations,” *Transportation Science*, 24(3), pp. 193–204.

Howard, R. A. and J. E. Matheson (1972): “Risk-sensitive Markov decision processes,” *Management Science*, 18(7), pp. 356–369.

Keeney, R. L. and H. Raiffa (1976): *Decisions with multiple objectives: preferences and value tradeoffs*, John Wiley & Sons, New York.

Kirkwood, C. W. (2004): “Approximating risk aversion in decision analysis applications,” *Decision Analysis*, 1(1), pp. 55–72.

Lancaster, J. (2003): “The financial risk of airline revenue management,” *Journal of Revenue and Pricing Management*, 2(2), pp. 158–165.

Lautenbacher, C. J. and S. J. Stidham (1999): “The underlying Markov decision process in the single-leg airline yield management problem,” *Transportation Science*, 33(2), pp. 136–146.

Lindenmeier, J. and D. K. Tscheulin (2005): “Kundenzufriedenheitsrelevante Effekte der Überbuchung im Rahmen des Revenue-Managements,” in: G. Fandel and H. B. von Portatius (eds.), *Zeitschrift für Betriebswirtschaft - Special Issue*, vol. 1, 1 edition, Gabler, Wiesbaden, pp. 101–123.

Littlewood, K. (1972): “Forecasting and control of passenger bookings,” *AGIFORS Symposium Proceedings*, pp. 95–117.

Mitra, D. and Q. Wang (2005): “Stochastic traffic engineering for demand uncertainty and risk-aware network revenue management,” *IEEE/ACM Transactions on Networking*, 13(2), pp. 221–233.

Pratt, J. W. (1964): “Risk aversion in the small and in the large,” *Econometrica*, 32(1/2), pp. 122–136.

Robinson, L. W. (1995): “Optimal and approximate control policies for airline booking with sequential nonmonotonic fare classes,” *Operations Research*, 43(2), pp. 252–263.

Subramanian, J., C. J. Lautenbacher, and S. Stidham (1999): “Airline yield management with overbooking, cancellations, and no-shows,” *Transportation Science*, 33(2), pp. 147–167.

Talluri, K. T. and G. J. van Ryzin (2004): *The theory and practice of revenue management*, 1 edition, Kluwer Academic Publishers, Boston.

Weatherford, L. R. (2004): "EMSR versus EMSU: revenue or utility," *Journal of Revenue and Pricing Management*, 3(3), pp. 274–284.

Weatherford, L. R. and S. E. Bodily (1992): "A taxonomy and research overview of perishable asset revenue management: yield management, overbooking, and pricing," *Operations Research*, 40(5), pp. 831–844.

Wirtz, J., S. E. Kimes, J. H. P. Theng, and P. Patterson (2003): "Revenue management: resolving potential customer conflicts," *Journal of Revenue and Pricing Management*, 2(3), pp. 216–226.

Wollmer, R. D. (1992): "An airline management model for a single leg route when lower fare classes book first," *Operations Research*, 40(1), pp. 26–37.

Visual Ontology Modeling for Electronic Markets

Saartje Brockmans[1], Andreas Geyer-Schulz[2], Pascal Hitzler[1], and Rudi Studer[1]

[1] Institute of Applied Informatics and Formal Description Methods,
Universität Karlsruhe (TH)
{brockmans,hitzler,studer}@aifb.uni-karlsruhe.de

[2] Institute of Information Systems and Management,
Universität Karlsruhe (TH)
andreas.geyer-schulz@em.uni-karlsruhe.de

Summary. Much research has been conducted on the applicability of semantic technologies to electronic markets. Semantic technologies promise to boost the interoperability of data, intelligent information management, and automated service management and service composition. This paper introduces our recent work on a visual, UML2-based notation to simplify formal descriptions with ontologies. Visual modeling reduces structural as well as typographical errors. Moreover, using UML2 improves acceptance in industry and accessibility by utilizing standard methods and tools. The recommended visual notation facilitates ontology engineering, especially for people who are familiar with UML2 rather than logical notation. Different application areas, including product descriptions for electronic markets or alignments of inter-organizational business process interfaces, benefit from the advantages of visual modeling of formal descriptions. In this paper, we define a metamodel for ontologies built on the Meta Object Facility, which is a model-driven integration framework for defining, manipulating and integrating both metadata and data in a platform-independent way. Based on this, we present a UML profile that determines a visual UML2 syntax to model ontologies. The UML profile is based on the UML2 class diagram notation. This notation is suited for ontology models since it is designed to describe the types of objects in a system and the various kinds of static relationships that exist among them. After presenting the metamodel and the UML profile, we examine a real-life application scenario in which ontologies are used as a solution for flexible product descriptions in the automotive industry.

1 Introduction

Semantic technologies are being heavily researched in relation to electronic markets. Among other things, they promise to achieve improvements in data interoperability and intelligent management of information. Several applications have already demonstrated the usability of ontologies for flexible product descriptions in markets, where knowledge has to be shared between various organizations and departments. An example of this is presented in a real-life project in the automotive industry. This particular setting illustrates the usefulness of ontologies in improving business processes and the interaction of different information sources very well Schnurr and Angele (2005). In this project, the configuration process of built-on-demand cars as well as test cars is enhanced by a guiding software assistant, which aids the engineer in taking the necessary dependencies into account and provides the potential expert contact information. The system also provides explanations so that the engineer can understand and validate the system´s decisions. Next to this use of ontologies for product descriptions, ontologies are also used as a means to integrate different information sources in a common language and thus to provide a unified view on the different data sources from the various suppliers. This project furnishes a concrete example from which similar usages of ontologies in markets can be imagined.

Unfortunately, the manual creation of ontologies is a labor-intensive, expensive, often difficult and – especially without proper tool support – an error-prone task. Visual syntaxes have been shown to bring many benefits that simplify conceptual modeling (Issing and Klimsa (2002)). Their usefulness has been demonstrated in practice: Visual modeling paradigms such as the Entity Relationship (ER) (Chen (1976)) model or the Unified Modeling Language (UML) (Fowler (2004)) are frequently used for the purpose of conceptual modeling. Nor surprisingly, the necessity of a visual syntax for knowledge representation languages has often been argued for in the past (Gaines (1991); Kremer (1998)). Particular representation formalisms such as conceptual graphs (Sowa (1992)) or Topic Maps (ISO/IEC (1999)) are based on well-defined graphical notations. Visual modeling decreases syntactic and semantic errors and increases readability. It makes the modeling and use of ontologies much simpler and faster, especially if the tools are user-friendly and appropriate modeling languages are applied. Therefore, we have developed a metamodel built on the Meta Object Facility (MOF) (OMG (2002)), a model-driven integration framework for defining, manipulating as well as integrating metadata and data in a platform-independent way. The metamodel, called Ontology Definition Metamodel (ODM), defines how ontologies are modeled, with specific focus on the standardized OWL DL language (Dean and Schreiber (2003)). In addition, we have developed a UML profile for the purpose of visual modeling, based on the UML class diagram notation (Fowler (2004)).

OWL DL ontologies do not include rules. However, an ontology without rules describes only relationships between concepts like "a part is a part of another part", "a part is connected to another part", etc. More complex relationships have to be described by rules. It is this more complex knowledge that has to be captured by the ontology to help, for instance, in configuring the test cars of the example scenario mentioned earlier. Rule extensions are currently being heavily discussed (W3C) and one of the most prominent proposals for an extension of OWL DL with rules is the Semantic Web Rule Language (SWRL) (Horrocks et al. (2004)). We extended the basic Ontology Definition Metamodel (supporting only OWL DL) and the corresponding UML profile to support SWRL (Brockmans et al. (2006b)).

This paper first describes a solution for visually modeling ontologies in Section 2. This solution consists of an Ontology Definition Metamodel as well as a visual syntax for modeling ontologies. After that, Section 3 describes an application scenario showing the results of our proposed solution. Finally, we make some concluding remarks and briefly discuss future work.

2 Visual Ontology Modeling

In 2004, the World Wide Web Consortium (W3C) finished its standardization work on the Web Ontology Language (OWL), thus laying the foundations for the widespread use of ontologies in business. Three variants of OWL have been defined: OWL Full, OWL DL (OWL Description Logic) and OWL Lite. With OWL DL, relevant concepts of the application domain, their properties and instances can be defined, and inferencing problems are decidable. OWL Lite has an even simpler syntax and decidable inference problems while OWL Full offers a more elaborate syntax but is undecidable. Since OWL DL is decidable and the syntax is sufficient, this variant is appropriate for our work.

In the meantime, rule extensions for OWL have been heavily discussed (W3C). Just recently, the W3C chartered a working group for the definition of a Rule Interchange Format (W3C (2005)). One of the most prominent proposals for an extension of OWL DL with rules is the Semantic Web Rule Language (SWRL) (Horrocks et al. (2004)). A high-level abstract syntax is provided that extends the OWL abstract syntax described in the OWL Semantics and Abstract Syntax document (Patel-Schneider et al. (2004)). An extension of the OWL model-theoretic semantics provides a formal meaning for SWRL ontologies.

As we explained before, our solution for the visual modeling of ontologies consists of a metamodel and a UML profile. As a foundation, we start by giving a short introduction to UML in Section 2.1. In Section 2.2, we describe the metamodel. This metamodel is based on OWL DL and SWRL, and defines how ontologies in these languages may be modeled. The visual (UML) syntax describes a UML profile, so it provides the engineer with a visual language based on UML to model the ontologies. This UML profile is explained in Section 2.3.

2.1 Unified Modeling Language

The Unified Modeling Language (UML) is a family of graphical notations, backed by a single metamodel, that help to express domain models. UML defines a notation and a metamodel. The notation is the graphical syntax of the modeling language. The metamodel is a model that defines the concepts of the language. UML is a well-established and very popular language standardized and driven by the Object Management Group.

One of the characteristics of UML2, which is an extension of UML, is that it is independent of the methodology that is used for analysis and design. Regardless of the methodology that you employ, you can use UML2 to express the results. And, using other standards, you can transfer your UML2 model from one specific tool into a repository, or into another specific tool for refinement or the next step in your chosen process. These are the main characteristics responsible for the language's widespread industry support.

UML2 defines thirteen types of diagrams; of these, the Class Diagram is the most interesting for our work. A class diagram gives an overview of a domain by showing its concepts and their attributes as well as the relationships among them. They also display which elements interact and show constraints on the interactions. But they do not illustrate what happens when an interaction takes place.

We use UML class diagrams as a means for the visual modeling of product descriptions using ontologies. We can benefit e.g. from reusing existing tools, and UML provides an agreed-upon format covering a broad community. There is a far greater likelihood that a user engaged in modeling product descriptions will be familiar with UML than with logic used to manually write ontologies, such as OWL DL.

2.2 An Ontology Definiton Metamodel

The Ontology Definition Metamodel defines a metamodel for ontologies. This metamodel is built on the Meta Object Facility OMG (2002), a model-driven integration framework for defining, manipulating and integrating both metadata and data in a platform-independent way. It provides a framework capable of supporting any kind of metadata and allows new kinds to be added as required. In short, it allows the definition of modeling languages.

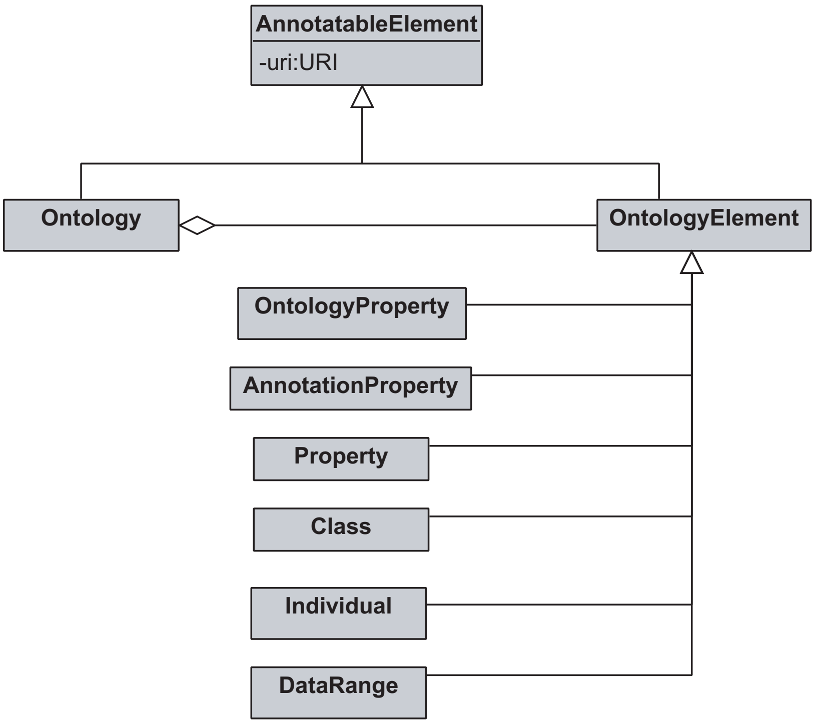


Figure 1: The main OWL DL-part of the Ontology Definition Metamodel

We have defined an Ontology Definition Metamodel for OWL DL extended with SWRL (Brockmans et al. (2006b)). A metamodel for a language that allows the definition of ontologies naturally follows from the modeling primitives offered by the ontology language. Naturally, our proposed metamodel has a one-to-one mapping to the abstract syntax of the two languages and thereby to their formal semantics. It primarily uses basic well-known concepts from UML2, including classes, associations and multiplicities, but also employs some extensions, such as stereotypes. Additionally, we have augmented the metamodel with constraints specifying invariants that have to be fulfilled by all models that instantiate the metamodel. These constraints are expressed in the Object Constraint Language (Warmer and Kleppe (2004)). Brockmans et al. (2004), Loeffler (2004) and Brockmans et al. (2006b) provide the sets of OCL constraints for the OWL DL-part and the SWRL-part of the metamodel.

The formal semantics of OWL is derived from Description Logics (Baader et al. (2003)), an extensively researched Knowledge Representation formalism. Hence, most primitives offered by OWL can also be found in Description Logics. Figure 1 shows an excerpt of the OWL DL-part of the Ontology Definition Metamodel. Users familiar with UML can, among other things, see that every element of an ontology, e.g. *Property* or *Datatype*, is an *OntologyElement* and hence a member of an *Ontology*. Moreover, one can discern from the model that an *Ontology* and every *OntologyElement* can be annotated; these two items are therefore generalized in a concept called *AnnotatableElement*. Figure 2 shows the main excerpt of the SWRL-part of the Ontology Definiton Metamodel. It is apparent from the model that a *Rule* consists of an *Antecedent* and a *Consequent*, both consisting of a set of *Atom*s which may or may not be empty. The *Atom*s of the *Antecedent* and the *Consequent* consist of *PredicateSymbol*s and *Term*s, which can have the form of an *Individual*, a *Variable* or a *DataValue*. For a detailed description of the full metamodel, we refer to Brockmans et al. (2006b, 2004); Loeffler (2004).

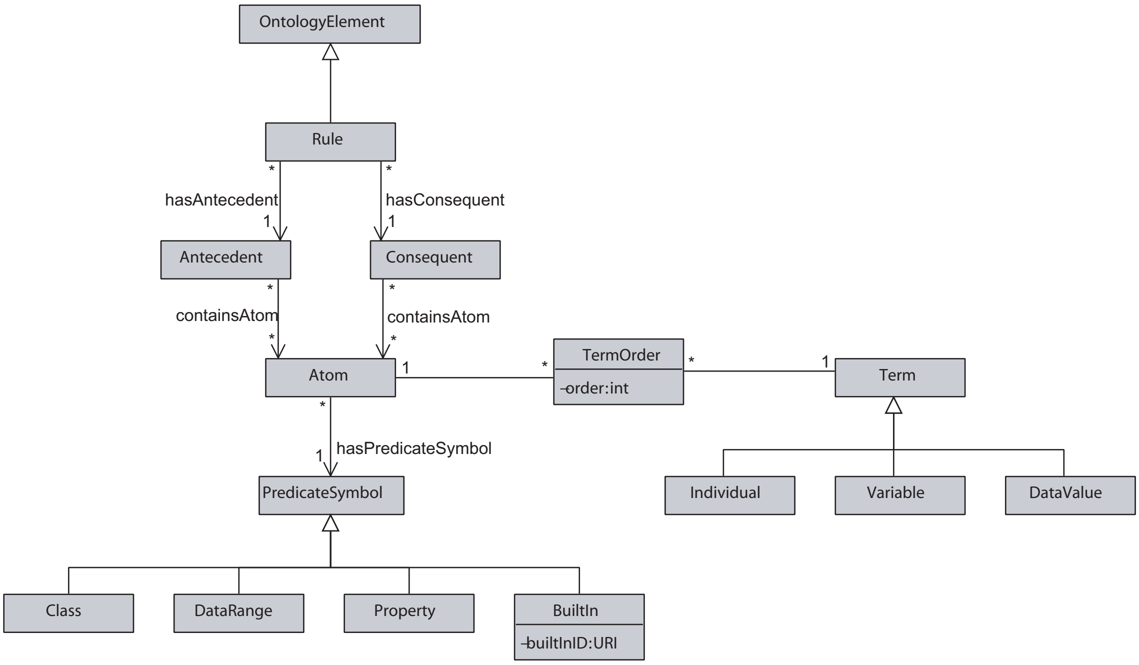


Figure 2: The main SWRL-part of the Ontology Definition Metamodel

2.3 A UML Profile

We require a UML profile in order to determine a visual UML syntax for modeling ontologies. We provide a UML profile that captures the intuitions behind UML as well as OWL and SWRL, with a maximal reuse of UML, OWL and SWRL features. Since the UML profile mechanism supports a restricted form of metamodeling, our proposal contains a set of extensions and constraints to the UML2 metamodel. This tailors UML such that models instantiating the Ontology Definition Metamodel can be defined. Our UML profile has a basic mapping, from class to class, property to n-ary association, individual to object and property filler to object association. Extensions to UML2 consist of custom UML-stereotypes, which usually carry the name of the corresponding OWL or SWRL language element, and dependencies. Figures 5-35 in Appendix A provide a list of mappings from OWL and SWRL to UML syntax, using examples from one of the most well-known ontologies, the so-called wine ontology.

Due to this easy-to-use creation of ontologies, we expect businesses to make use of ontologies to a higher degree in the future; an immediate benefit is apparent in the time saved by ensuring correct flexible configurations.

3 Ontologies in the Automotive Industry

This section describes a project in the automotive industry where ontologies are used for representing and sharing knowledge. This approach enhances business processes for flexible configurations of built-on-demand cars and facilitates the integration of life data into this optimization process. Throughout the discussion, we provide excerpts of the visual ontology model built using our UML profile described in Section 2.

The company in our real-life application scenario produces cars. Nowadays, many cars are built on demand, which means that clients can buy cars which are configured to fit their personal needs and wishes. Moreover, the company has a fleet of cars which are continually reconfigured to test all kinds of new parts and new combinations. When new car parts enter the market, new dependencies arise which are often only known by a handful of human experts. This scenario demands a lot of communication both among and between different suppliers and manufacturers. Not surprisingly, this pattern results in many errors. When the built-on-demand cars are ready, they often appear flawed or in some cases do not even work. Considering that configuring a test car takes about two weeks, this clearly represents a major loss of time for a company. The objective therefore is to shorten the time-to-market via reducing both the number of errors and the amount of communication by making the understanding of product descriptions and their dependencies exploitable by computers.

The process of flexible configuration of the cars is enhanced by a guiding software assistant that helps the engineer to take the necessary dependencies into account and provides the potential expert contact information. The system also makes explanations available so that the engineer can understand and validate the system's decisions. All this information is stored in an ontology, which we will call the automotive ontology for now.

This ontology has to be created and written in a logical language. But chances are high that the engineers who have to create these product descriptions will not be familiar with writing logic. However, many engineers are familiar with UML. Moreover, visual modeling has proven itself to be less error-prone, as discussed in Section 1. And this is where our solution from Section 2 comes in.

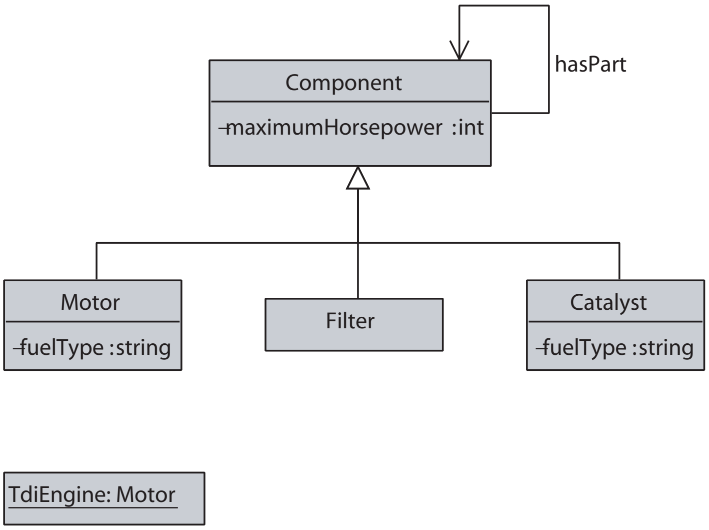


Figure 3: Small excerpt of the automotive industry ontology

Figure 3 shows part of the ontology as it looks when modeled by an engineer. A basic concept *Component* with a relationship *hasPart* to another *Component* and an attribute *maximumHorsepower* is defined. A *Motor*, a *Filter* and a *Catalyst* are all defined as special types of *Component*. *Motor* and *Catalyst* both have the attribute *fuelType*. An instance *TdiEngine* is defined as a specific kind of *motor*.

One of the rules defining relationships between different parts states that the filter installed in a catalyst must be able to filter the motor's fuel. Figure 4 shows the visual representation of this rule using the UML profile.

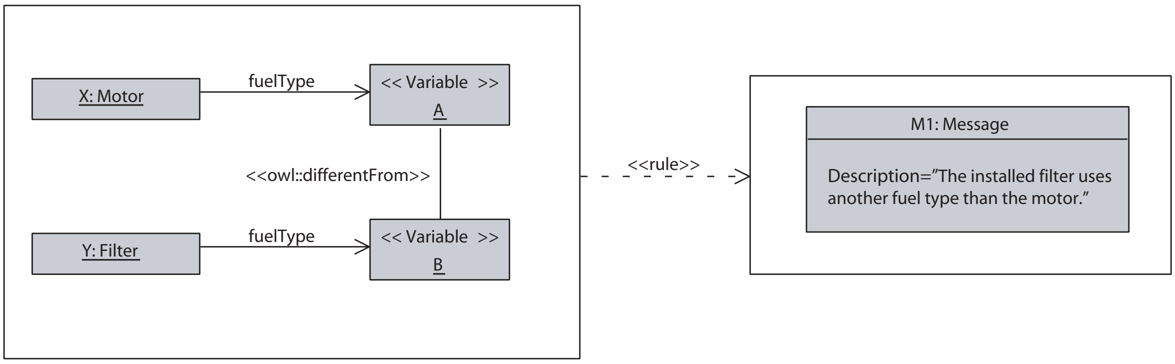


Figure 4: Message("The installed filter uses another fuel type than the motor") ← $\text{Motor}(X) \wedge \text{fuelType}(X, A) \wedge \text{Filter}(Y) \wedge \text{fuelType}(Y, B) \wedge \text{differentFrom}(A, B)$

All visual models are created using an existing UML tool supporting UML profiles. The tool is extended to automatically translate the models into OWL DL and SWRL so that it can be used by the guiding software assistant.

In addition to be being deployed for product descriptions, ontologies also serve to integrate the different information sources from the different departments and manufacturers into the common automotive ontology. This mapping from the most recent additional product information from the manufacturers to the existing automotive ontology happens at real-time when the engineer uses the guiding software assistant. This assures that the ontology used during the configuration of the test cars, is always up-to-date. Clearly, the different information sources sometimes contain redundant or even inconsistent information. These problems are handled with rules that can be visually modeled in just the same way as the rule shown in Figure 4.

4 Conclusion and Outlook

In this paper, we presented a solution to make the advantages of ontologies more accessible for engineers in markets who are e.g. describing products. A MOF metamodel for the Web Ontology Language OWL and the Semantic Web Rule Language SWRL is provided for this purpose. The validity of instances of this metamodel is ensured through OCL constraints. In connection with this metamodel, we also provided a UML profile, which allows engineers to model ontologies in a more intuitive visual way instead of writing logic. We illustrated our solution via a real-life application scenario which underscores the need for ontologies in markets. By exploring this example, one can easily imagine other concrete needs in markets. We describe another application area in Brockmans et al. (2006a), where ontologies are visually modeled for semantic alignment of business processes. For the further development and the standardization of the Ontology Definition Metamodel and the UML profile, we are working closely together with the corresponding working group at the Object Management Group (OMG) (IBM and Sandpiper Software Inc. (2005)). Moreover, to support the different rule languages that different companies use, we will provide additional metamodels and mappings between the different rule formalisms.

Acknowledgements

We would like to thank our colleagues, especially Peter Haase, for their valuable input and feedback.

References

Baader, F., D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider (eds.) (2003): *The Description Logic Handbook*, Cambridge University Press.

Brockmans, S., M. Ehrig, A. Koschmider, A. Oberweis, and R. Studer (2006a): "Semantic Alignment of Business Processes," in: *8th International Conference on Enterprise Information Systems*.

Brockmans, S., P. Haase, P. Hitzler, and R. Studer (2006b): "A Metamodel and UML Profile for Rule-extended OWL DL Ontologies," in: *3rd Annual European Semantic Web Conference*, Springer, Budva, Montenegro.

Brockmans, S., R. Volz, A. Eberhart, and P. Loeffler (2004): "Visual Modeling of OWL DL Ontologies using UML," in: S. A. McIlraith, D. Plexousakis, and F. van Harmelen (eds.), *The Semantic Web - ISWC 2004*, *Lecture Notes in Computer Science*, vol. 3298, Third International Semantic Web Conference, Springer, Hiroshima, Japan, pp. 198–213.

Chen, P. (1976): "The Entity-Relationship Model–Toward a Unified View of Data," *ACM Transactions on Database Systems*, 1(1), pp. 9–36.

Dean, M. and G. Schreiber (2003): "Web Ontology Language (OWL) Reference Version 1.0," working paper, World Wide Web Consortium (W3C), internet: `http://www.w3.org/TR/owl-ref/`.

Fowler, M. (2004): *UML Distilled Third Edition*, Addison-Wesley.

Gaines, B. R. (1991): "An Interactive Visual Language for Term Subsumption Languages," in: J. Mylopoulos and R. Reiter (eds.), *Proceedings of 12th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, Sydney, Australia, pp. 817–823.

Horrocks, I., P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean (2004): *SWRL: A Semantic Web Rule Language Combining OWL and RuleML*, World Wide Web Consortium, internet: http://www.w3.org/Submission/2004/SUBM-SWRL-20040521/.

IBM and Sandpiper Software Inc. (2005): "Ontology Definition Metamodel," working paper, OMG, internet: http://www.omg.org/docs/ad/05-08-01.pdf.

ISO/IEC (1999): "Topic Maps: Information Technology – Document Description and Processing Languages," ISO/IEC standard 13250:2000, internet: http://www.y12.doe.gov/sgml/sc34/document/0129.pdf.

Issing, L. J. and P. Klimsa (eds.) (2002): *Wissenserwerb mit Texten, Bildern und Diagrammen*, third edition, Belz, PVU, Weinheim, pp. 65–81.

Kremer, R. (1998): "Visual Languages for Knowledge Representation," in: *Proc. of 11th Workshop on Knowledge Acquisition, Modeling and Management (KAW'98)*, Morgan Kaufmann, Voyager Inn, Banff, Alberta, Canada.

Loeffler, P. (2004): *UML zur Visuellen Modellierung von OWL DL*, Master's thesis, Institute AIFB, Universität Karlsruhe.

OMG (2002): "Meta Object Facility (MOF) Specification," working paper, Object Management Group (OMG), internet: http://www.omg.org/docs/formal/02-04-03.pdf.

Patel-Schneider, P. F., P. Hayes, and I. Horrocks (2004): "OWL Web Ontology Language Semantics and Abstract Syntax," working paper, World Wide Web

Consortium, recommendation. http://www.w3.org/TR/2004/REC-owl-semantics-20040210/.

Schnurr, H.-P. and J. Angele (2005): "Do Not Use This Gear with a Switching Lever! Automotive Industry Experience with Semantic Guides," in: Y. Gil, E. Motta, V. R. Benjamins, and M. Musen (eds.), *The Semantic Web – ISWC 2005*, number 3729 in Lecture Notes in Computer Science, 4th International Semantic Web Conference, Springer-Verlag, Galway, Ireland, pp. 1029–1040.

Sowa, J. F. (1992): "Conceptual Graphs Summary," in: P. Eklund, T. Nagle, J. Nagle, and L. Gerholz (eds.), *Conceptual Structures: Current Research and Practice*, pp. 3–52.

W3C (2005): *Accepted Papers of the W3C Workshop on Rule Languages for Interoperability*, Washington, DC, USA, http://www.w3.org/2004/12/rules-ws/accepted.

W3C (2005): "Rule Interchange Format Working Group Charter (RIF)," `http://www.w3.org/2005/rules/wg/charter`.

Warmer, J. and A. Kleppe (2004): *Object Constraint Language 2.0*, MITP Verlag.

A - Mappings from OWL and SWRL to UML syntax

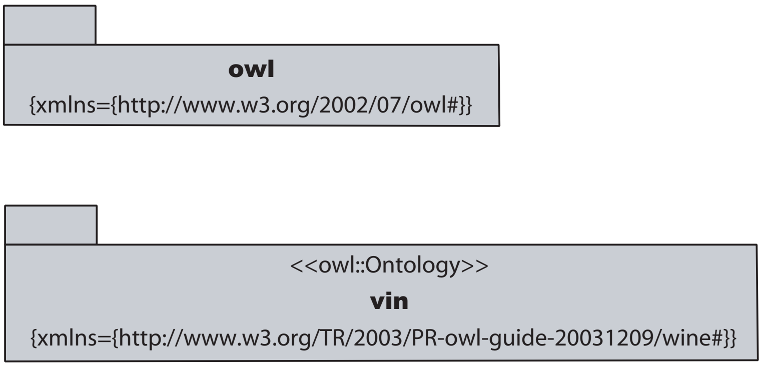

owl
{xmlns={http://www.w3.org/2002/07/owl#}}
<<owl::Ontology>>
vin
{xmlns={http://www.w3.org/TR/2003/PR-owl-guide-20031209/wine#}}

Figure 5: Namespace(owl=http://www.w3. org/2002/07/owl#) Namespace(vin=http://www.w3.org/TR/ 2003/PR-owl-guide-20031209/wine#) Ontology(http://www.w3.org/TR/2003/ PR-owl-guide-20031209/wine)

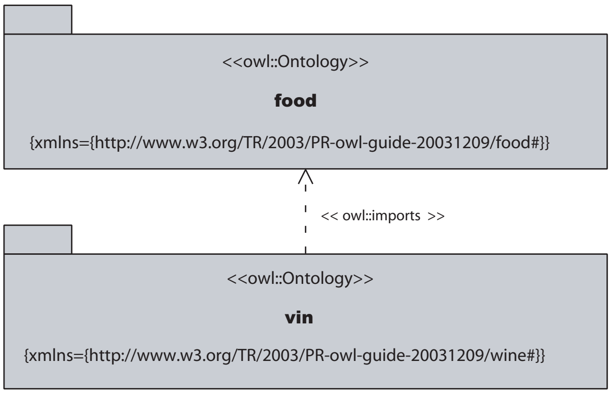

<<owl::Ontology>>
food
{xmlns={http://www.w3.org/TR/2003/PR-owl-guide-20031209/food#}}
<< owl::imports >>
<<owl::Ontology>>
vin
{xmlns={http://www.w3.org/TR/2003/PR-owl-guide-20031209/wine#}}

Figure 6: Annotation(owl:imports food)

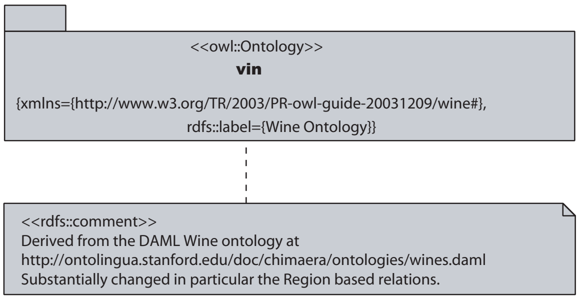

<<owl::Ontology>>
vin
{xmlns={http://www.w3.org/TR/2003/PR-owl-guide-20031209/wine#},
rdfs::label={Wine Ontology}}
<<rdfs::comment>>
Derived from the DAML Wine ontology at
http://ontolingua.stanford.edu/doc/chimaera/ontologies/wines.daml
Substantially changed in particular the Region based relations.

Figure 7: Annotation(rdfs:comment "Derived from the DAML Wine ontology at http://ontolingua.stanford.edu/doc/ chimaera/ontologies/wines.daml Substantially changed, in particular the Region based relations.") Annotation(rdfs:label "Wine Ontology")

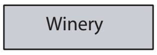


Figure 8: Class(Winery)

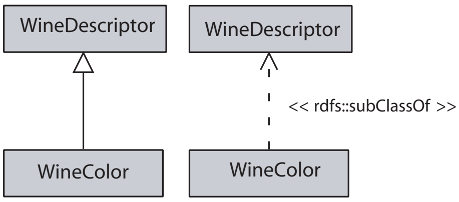


Figure 9: Class(WineColor partial WineDescriptor) (two alternative notations)

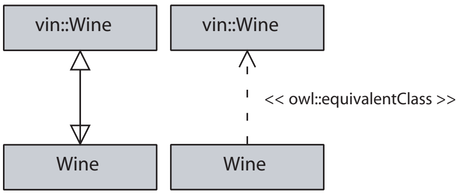


Figure 10: EquivalentClasses(Wine vin:Wine) (two alternative notations)

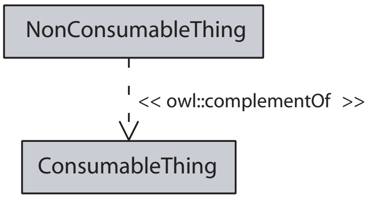


Figure 11: Class(NonConsumableThing complete complementOf(ConsumableThing))

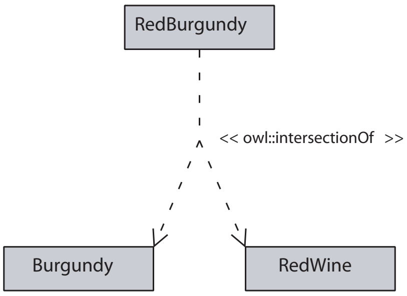


Figure 12: Class(RedBurgundy complete intersectionOf (Burgundy RedWine))

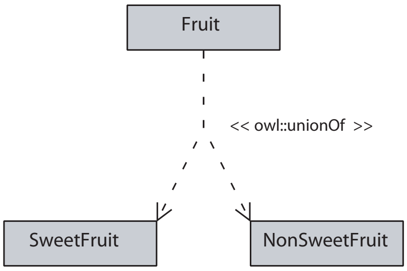


Figure 13: Class(Fruit complete unionOf(NonSweetFruit SweetFruit))

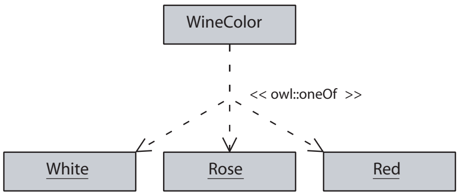


Figure 14: EnumeratedClass(WineColor White Rose Red)

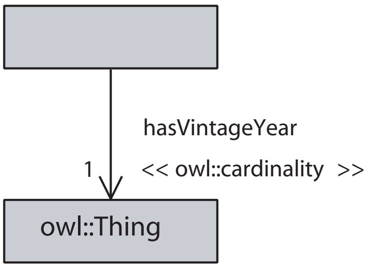


Figure 15: restriction(hasVintageYear cardinality(1))

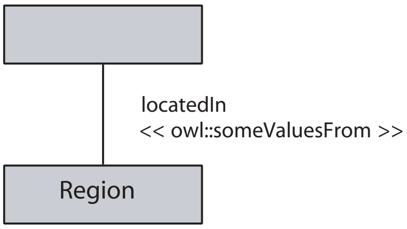


Figure 16: restriction(locatedIn someValuesFrom(Region))

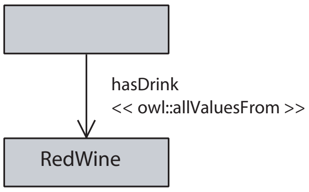


Figure 17: restriction(hasDrink allValuesFrom RedWine)

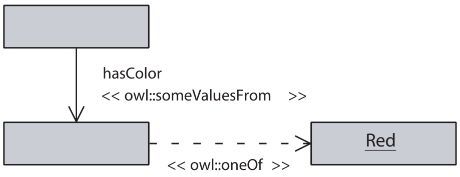


Figure 18: restriction(hasColor hasValue(Red))

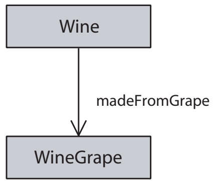


Figure 19: ObjectProperty(madeFromGrape domain(Wine) range(WineGrape))

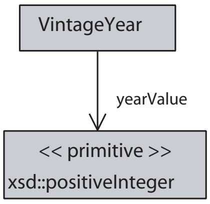


Figure 20: DatatypeProperty(yearValue domain(VintageYear) range(xsd: positiveInteger))

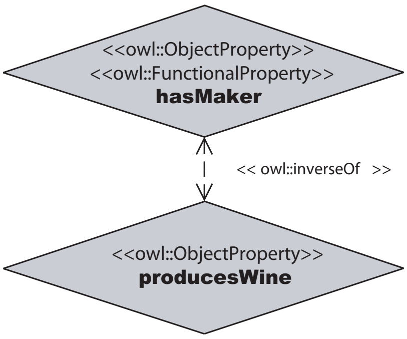


Figure 21: ObjectProperty(hasMaker Functional), ObjectProperty(producesWine inverseOf(hasMaker)))

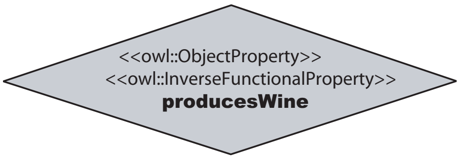


Figure 22: ObjectProperty(producesWine InverseFunctional)

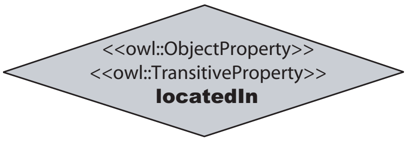


Figure 23: ObjectProperty(locatedIn Transitive)

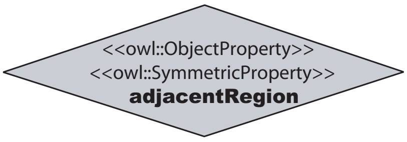


Figure 24: ObjectProperty(adjacentRegion Symmetric)

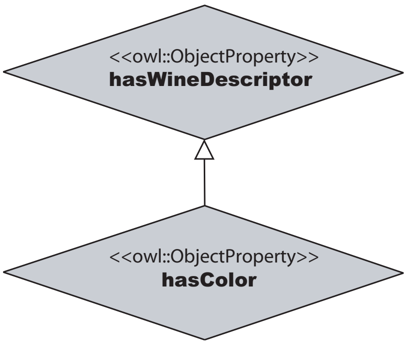


Figure 25: SubPropertyOf(hasColor hasWineDescriptor)

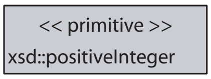


Figure 26: xsd:positiveInteger

CabernetSauvignonGrape:Winegrape

Figure 27: Individual(CabernetSauvignonGrape type(WineGrape))

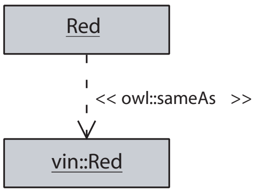


Figure 28: SameIndividual(Red vin:Red)

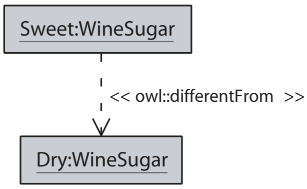


Figure 29: DifferentIndividuals(Sweet Dry)

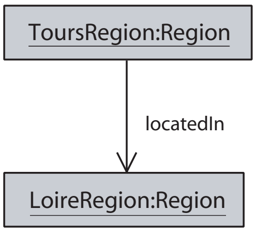


Figure 30: Individual(ToursRegion type(Region) value(locatedIn LoireRegion))

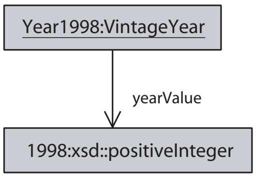


Figure 31: Individual(Year1998 type(VintageYear) value(yearValue "1998" xsd:positiveInteger))

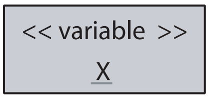


Figure 32: Variable x

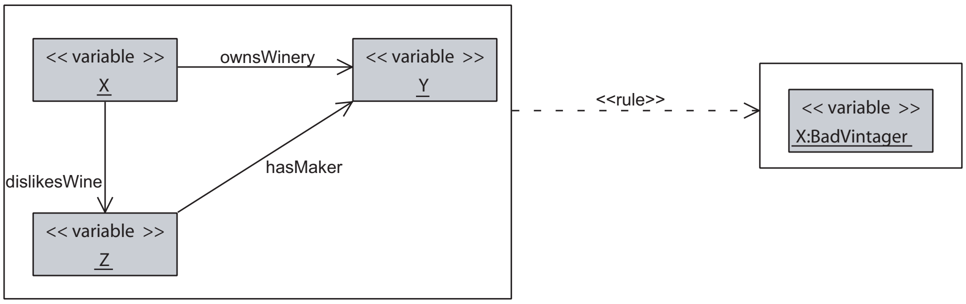


Figure 33: BadVintager$(x) \leftarrow$ ownsWinery$(x, y) \wedge$ dislikesWine$(x, z) \wedge$ hasMaker(z, y)

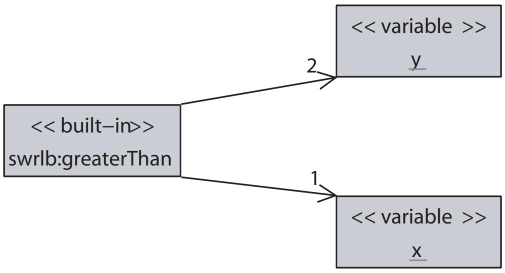


Figure 34: builtIn(greaterThan x y)

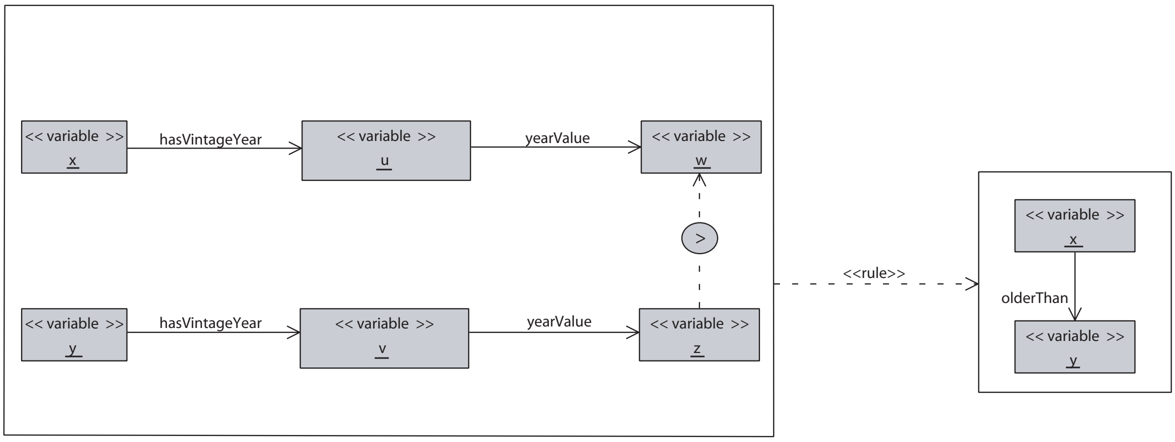


Figure 35: olderThan(x, y) ← hasVintageYear(x, u) $\wedge$ hasVintageYear(y, v) $\wedge$ yearValue(u, w) $\wedge$ yearValue(v, z) $\wedge$ swrlb:greaterThan(z, w)

Using Network Analysis for Fraud Detection in Electronic Markets

Michael Blume[1], Christof Weinhardt[1], and Detlef Seese[2]

[1] Institute of Information Systems and Management,
Universität Karlsruhe (TH) {blume,weinhardt}@iw.uni-karlsruhe.de

[2] Institute of Applied Informatics and Formal Description Methods (AIFB)
Universität Karlsruhe (TH)
seese@aifb.uni-karlsruhe.de

Summary. Electronic markets are used for resource allocation within a wide spectrum of application domains, ranging from securities on stock exchanges to consumer goods on C2C online auctions. High benefits often tempt participants to open a second account for additional transactions. These benefits differ from platform to platform and depend on the intrinsic incentives. For instance on platforms with a reputation system, a high number of transactions increases a participant's reputation if there are no transaction costs. Thus, trading with himself may raise a participant's status to a better level (experienced user, frequent buyer, etc.). In online auctions, participants bid for their own good because they want to start a bidding war or do not estimate the prices offered so far as acceptable. On a common stock exchange; it can be advantageous for participants planning hidden takeovers to transfer money to another account. Another motivation for transferring shares from one account to another can be tax advantages in another country. Generally, all of these platforms explicitly prohibit using an account for malicious intent (shilling and sybil attacks).

In most cases, no reliable control mechanism exists to prevent participants from having multiple accounts. Therefore, the market operator must be able to identify these accounts or malicious transactions by analysing the market transaction data. Common automated approaches calculate indices based on transaction data. To show all relevant cases with a rule-based system, the threshold has to be so low that many cases end up being reported, that are not fraudulent (false positives). This stems from computing indices which only take transaction data into consideration without looking at the context of the actor within the market topology. The reliable isolation of serious cases for further investigation increases efficiency and saves time for the market operator.

This paper suggests a network analysis approach combined with interactive visualization to improve the automatic search for malicious accounts. In the first step, an individual node index similar to a network centrality measure is calculated. In the second step, suspicious nodes are displayed in their topological context, together with detailed background information about their trade history. The visual feedback is very important, as supervising is still done by humans.

Furthermore, we suggest implementing a machine-learning algorithm to improve the results over time. If the system detects a new suspicious case, the user will further investigate it, or, if it is obviously not a case of fraude, ignore it. This feedback helps to rate cases by importance and thus raises the probability of reporting true fraud. The resulting tool is dedicated for the application by exchange supervisory authorities of electronic online platforms.

1 Introduction

Trust and reliability are key factors for people considering joining a market. When it is possible to commit fraud with impunity the market will attract free riders and lose honest customers. Any serious market platform therefore has to enforce the rules. On common stock exchanges in Germany, for example, the legislature has already implemented a supervisory institution (HÜSt) by law (BRD, 2002, 1998). In the United States, meanwhile the U.S. Securities and Exchange Comission has charged the NYSE itself to do the investor protection and within that program market surveilance. Companies running online platforms share the same interest in having a clean market. Bad press due to fraud can cause users to migrate to competitor platforms. The number of reported frauds in online auctions rises each year, and with more than 51000 reported cases in 2002, constitued the largest category of Internet-related complaints to the U.S. Federal Trade Commission (FTC) (Robertson Barrett, 2003; Library, 2003). The financial damage caused by fraud in online auctions in the United States amounted to $437 million in 2003 (Wahab, 2004), clearly there is a need for supervision.

Online market platforms are becoming very popular, recent user statistics from Yahoo, eBay and Amazon clearly indicate the impossibility of supervising every account manually. Since the supervisory board needs tools and heuristics to flag suspicious cases, many companies develop their own tools (Baquia, 2002; Jennifer Chasin, 2000). The complexity of the problem arises from the intricate relationship between all observable facts. A single transaction without context is extremely difficult to evaluate in terms of fraud. Even the totality of transactions from a single account may not provide enough information to respond to this question with certainty. However, studying all of the transactions (including the corresponding counterparty for each transaction) from various perspectives (e.g. time constraints or monetary incentives) may furnish enough information for a reliable indicator. Due to the complex level of relationship between the different entities, we suggest a network analysis approach.

The article is structured as follows: In the next section, we will introduce the platform and dataset used in our research. In Section 3, we will describe the indicators. The results, conclusion and outlook are presented in the last two sections.

2 Platform and Dataset

To test the indicators developed in this paper we used a dataset from the information forecasting market for the “Herbstmeister” soccer championship in Germany in 2005 on STOCCER[1] (Luckner et al., 2005). “Herbstmeister” refers to the autumn championship

[1] STOCCER on `http://www.stoccer.de`

played by the teams in Germany's premiere federal football league. (For more information about forecasting markets, please refer to (Forsythe et al., 1992; Spann and Skiera, 2003, 2004)). STOCCER is a information forecasting market where participants can trade their expectations regarding the outcome of the football championship. Each team is represented by a share, which can be traded on the platform. So the actual market price of a share reflects the expectations regarding the likelihood (averaged over all participants) that the corresponding team will win the championship.

The quality of the market depends on the number of participants and their knowledgeability (Ortner, 1996; Nelson et al., 2003). Our dataset consisted of 134 participants with exactly 4900 transactions. The advantage of information forecasting markets is that a small number of participants, relative to the usual number of participants in surveys is sufficient to achieve fairly accurate predictions. As an incentive to participate in the forecasting market the best account is usually rewarded with a prize. In the Herbstmeister-market, the prize was free-of-charge participation in the forecasting market for the FIFA World Cup 2006.

In this dataset, some accounts are known to have violated the rules. The problem with a real dataset is that there may be even more fraud cases that have not yet come to our attention. On other online markets (such as auction platforms) there are even more opportunities to commit fraud (c.f. Wahab, 2004); these will not be considered in this article.

3 Indicators

Malicious accounts on information forecasting markets have several properties. They do not all necessarily have to apply and the existence of any one of them does not mark an account as malicious with certainty. Nevertheless, the detection of many of them at a time increase the probability of having found a suspicious account. We therefore need to design indicators for each of these properties. Every indicator will be based on assumptions reflecting market rules or incentives.

Generally speaking, the indicator approach consists of a plug-in-based system in which several indicators are calculated independently. For every element (actors and transactions) the resulting probability of malicious tendencies is stored as a separate value. In the end, all indicators are added up to yield an overall score. Vertices passing a certain threshold are reported to the user.

The indicators used for the different aspects of fraud are derived from vague textual descriptions of typical cases. It is impossible to derive sharp parameterised values or clear borders to identify suspicious cases from these descriptions. In any case, publicly known fixed thresholds do not make sense; any fraudulent participant would simply avoid reaching them. Thresholds have to be calculated dynamically and thus be adapted to the actual market situation. The list of indicators may vary from platform to platform; the incentive structure and rules may differ, and the ways of bypassing the rules or committing fraud will change accordingly. In this article we will briefly introduce some of the main ideas and explain one of the main indicators in more detail. The indicators presented below are designed for information forecasting markets and are based on several assumptions.

Assumption 1. *Every participant has a limited amount of time.*

Assumption 2. *A normal account always tries to become richer.*

The first assumption is that every participant has a limited amount of time to spend on the platform and thus devote to an account. In cases of fraud, we therefore expect to find one main account (primary account) and one or several smaller accounts (secondary accounts) that support the primary account ("small" refers to the amount of time invested). The probability of a user having two accounts of equal size is even lower, if the platform provides an incentive according to the size or value of an account (i.e. the best players are rewarded). This assumption will influence several indicators, e.g. the activity indicator (see Section 3.5) and the primary-secondary indicator (see Section 3.3).

Assumption 2 reflects the oft-cited "rationality" of the participant. Knowing that there are rewards for having the best account, a rational user will strive to enrich his account. In a fraud case, a rich account (the primary account) will strive to enrich itself at the expense of the secondary account(s). This we call the primary-secondary indicator.

Assumption 3. *Only manual trading is possible.*

Assumption 4. *The costs for opening an account are lower (or zero) than the benefits (e.g. initial equipment/reputation).*

Thridly we assume that only manual trading is possible. There is no software bidding interface or API. The manual trading assumption is related to the first assumption that participants can only spend a certain amount of time on the platform. If they had software agents at their disposal, this assumption would be violated: they could effortlessly control any number of accounts.

The last assumption is that in case of stock exchanges there are less or no costs for opening an account and every account has an initial amount of money. This assumption reflects the situation on STOCCER but can be relaxed so that opening an account results in a positive value for the user (opening costs are lower than benefits).

To know which indicators to use, the incentives of the underlying platform have to be reviewed first. For the information forecasting market STOCCER the main incentive is having the best account in the end. The best account is determined by multiplying its assets (i.e. money and winning team shares) by 100 monetary units. Since every account starts with a certain amount of money, the easiest way of obtaining more is to create a second account and transfer its monies to the first account (Assumption 3).

Table 1: Indicators and motivations

Indicators	Motivation
Second Account/ Name Indicator	Creativity may be limited such that there are similarities between registration data (user names, e-mail addresses, etc).
Circular Trading Indicator	If two people always trade in the spread in contrary directions with more or less the same amount of shares, they are probably trying to transfer money. This can be detected by a high number of trades within a certain period of time in which one account ends up with a much better monetary result than we would expect on an average price.
Order Book Indicator	Transferring money from one account to another is facilitated if prices may be set freely. This is only possible if the order book is empty on at least one side. Therefore, if few accounts buy the whole side of an order book within a short period of time and continue trading on different prices, our indicator will alert us. This indicator will not be explained in detail in this article.
Primary-Secondary Indicator	If only the best account wins the price, a user will always try to have a strong primary account and one more weak secondary accounts.
Prominent Edge Indicator	Besides possibly conducting regular trades as a smokescreen, a secondary account will have one prominent link with one or more transactions in order to transfer money to the primary account.
Activity Indicator	Activity, such as when an account was created, a transaction was concluded or an offer was placed, is documented by time stamps. The multiple accounts are often used simultaneously, but sometimes consecutively. Another factor is that some sorts of transfers (money or share) are more easily done when the market is relatively calm. Finally, secondary accounts are usually created after the primary account and are used for few days only.

This process can be repeated over and over again. The money transfer itself may also be transacted in different ways (see Table 1).

For the purpose of describing the proposed indicators some definitions (following (Jungnickel, 2005)) will be introduced first. A market can be easily transformed into an evaluated graph $G = \{V, E, \omega\}$, where V is the set of vertices, $E \subseteq E^2$ is the set of edges, i.e. tuples $\{a, b\}$ with $a, b \in V$. ω is an evaluation of vertices and edges, i.e.

a mapping of vertices and edges to a weighted function (i.e. $(0,1)$). The weight of an entity is denoted by $\omega(a)$, $a \in V \cup E$. Actors can be transformed into vertices and each sell or buy transaction into a directed edge (arc) from the seller to the buyer. Further denotes $deg(v)$ with $v \in V$ the degree (number of in- and out going arcs) of a vertex.

3.1 Name Indicator

The second account may have similar or similar-sounding registration data (e.g. similar e-mail nicknames or domains). This may be entirely innocent: it is conceivable that a user could create an account, abandon it due to failure to understand the platform, and later open a new account. However, it is not allowed to purposefully transfer money between these accounts. Similar registration data can be easily found with phonetic search algorithms like soundex (c.f. Philips, 2000) or double metaphone (c.f. Russel, 1918).

3.2 Circular Trading Indicator

There are many reasons why two participants might have several buy and sell transactions for a certain share between each other. So why and how to look for circular trading? The key here is to detect participants doing this repeatedly in order to transfer a small amount of money each time. Therefore, one indicator distinguishing random trades from circular trading will be the consistent buying and selling of (more or less) the same quantity of shares. Users engaging in circular trading depend on uninterrupted periods of activity and must therefore exclude other participants. The easiest way to do so is to trade in times of low activity or in shares with less common appeal.

From a graph-theoretical perspective, we are looking for a strong bi-derictional edge in the graph generated from the transactions of one type of stock only (i.e. strong in terms of the amount of shares or transactions). This means we will have as many graphs as shares with identical vertices but different edges. The runtime complexity is $O(\sum_{k \in S} |E_k|)$ with S representing the number of shares traded on the market and E_k denoting the number of edges connected to node k.

3.3 Primary-Secondary Indicator

This indicator is calculated for every transaction (i.e. edge in the graph). The value of these accounts is compared, looking for instances of a secondary account transferring money or shares to a primary account. To introduce this measurement, we need the following definition:

Every edge has a start vertex ($e_i(1)$, first vertex of the i-th edge, seller) and an end vertex ($e_i(2)$, second vertex of the i-th edge, buyer). Let $f(v,w)$, $v, w \in V$ be a function comparing the value of two vertices representing two accounts. Then let $p(e)$, $e \in E$ symbolise the price of the transactions with $p_{avg}(e)$, $e \in E$ denoting the average price of the last c transactions (c constant) for the share traded in transaction e. Finally $FP_e(i)$ denotes the number of fraud-points of edge i.

```
$\forall i \in |E|:$
$if\ (((f(e_i(1), e_i(2)) < 0)\, and\, (p(e_i) < p_{avg}(e_i)))\, or$
  $((f(e_i(1), e_i(2)) > 0)\, and\, (p(e_i) > p_{avg}(e_i))))\, then$
    $FP_e(i) = FP_e(i) + 1$
```

Algorithm 1: Primary-Secondary Indicator

Algorithm 1 marks strong accounts buying low or selling high and has a runtime complexity of $O(|T|)$, T the set of transactions. The f function evaluates the size of an account. The difference in size is not the only relevant factor, however. Usually the secondary account has not traded a lot at this point. At the same time the primary account usually has many more and/or stronger ties to other accounts in the graph and has traded with better "established"/older participants than the secondary account. These aspects need to be incorporated into the evaluation function f.

Algorithm 2 shows a possible implementation of the function f. First we calculate the ratio of older counterparties to newer counterparties for each account compared. Then we evaluate the wealth, number of transactions and which of the two accounts was created first. All of these values are multiplied by their specific weight and added up to return the final value. This function is a good place to include network centrality measures (a good overview is given in (Brandes and Erlebach, 2005)). A central, well-established vertex is more likely to be a primary account while a peripheral vertex will more often be a secondary account. Degree centrality ($C(v) = \frac{deg(v)}{\sum_{w \in V} deg(w)}$ with $v \in V$) suits very well, since buy and sell transactions only imply a depth of one and will not influence distant vertices. In addition, betweeness centrality ($C(v) = \sum_{s,t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$ with $v \in V$, σ_{st} denoting all paths from s to t and $\sigma_{st}(v)$ denoting all paths between s and t going through v) is a good measure for indicating how well a vertex is embedded within a network. These measures can be multiplied by the overall weight to derive the embeddedness of the vertex.

3.4 Prominent Edge Indicator

If a second account has been created to enrich the primary account, it must somehow transfer money to the latter. Thus, if we look at the sum of all of the transactions, we will see that the majority is related to just one partner. This leads us to conclude that every vertex having just one prominent edge (summing all transactions with every neighbour together) and only few or comparably weak other edges, is suspicious. It is still not a strong indicator for fraud, however, because it may be the case that someone simply wanted to get rid of a larger share and sold it at a good price, or that only few accounts provide liquidity (and thus the main trading occurs between them and the rest). It may even be that it was simply the first transaction made by this account. In this case, it can only have one big transaction, compared to no prior activity!

Algorithm 3 is simple. We look in the set of edges of a single vertex whether the traded volume over a certain edge is larger than the weighted average. If this is true for more than one we can skip that vertex; otherwise we mark the corresponding edge and vertex as suspicious.

```
$f(v,w),\ v,w \in V$
  $E_v = \{e \in E | e \cap v \neq \emptyset\}$
  $E_w = \{e \in E | e \cap w \neq \emptyset\}$
  $age_v = 0$
  $\forall e \in E_v$:
    $if\ (age\_of(v) > age\_of(e \backslash v))\, then$
        $age_v + +$
  $age_w = 0$
  $\forall e \in E_w$:
    $if\ (age\_of(w) > age\_of(e \backslash w))\, then$
        $age_w + +$
  $\omega_{age} = \frac{age_v}{|E_v|} - \frac{age_w}{|E_w|}$
  $\omega_{wealth} = \frac{wealth(v)}{avg_{wealth}} - \frac{wealth(w)}{avg_{wealth}}$
  $\omega_{transaction} = \frac{getNum\_ofTransactions(v)}{avg_{transactions}} - \frac{getNum\_ofTransactions(w)}{avg_{transactions}}$
  $if\ (age\_of(v) > age\_of(w))\, then$
   $\omega_{firstCreated} = c_{firstCreated}$
  else
   $\omega_{firstCreated} = -c_{firstCreated}$
  $f(v,w) = \omega_{firstCreated} + \omega_{age} * c_{age} + \omega_{wealth} * c_{wealth} + \omega_{transaction} * c_{transaction}$
```

Algorithm 2: Example of an account evaluation function f

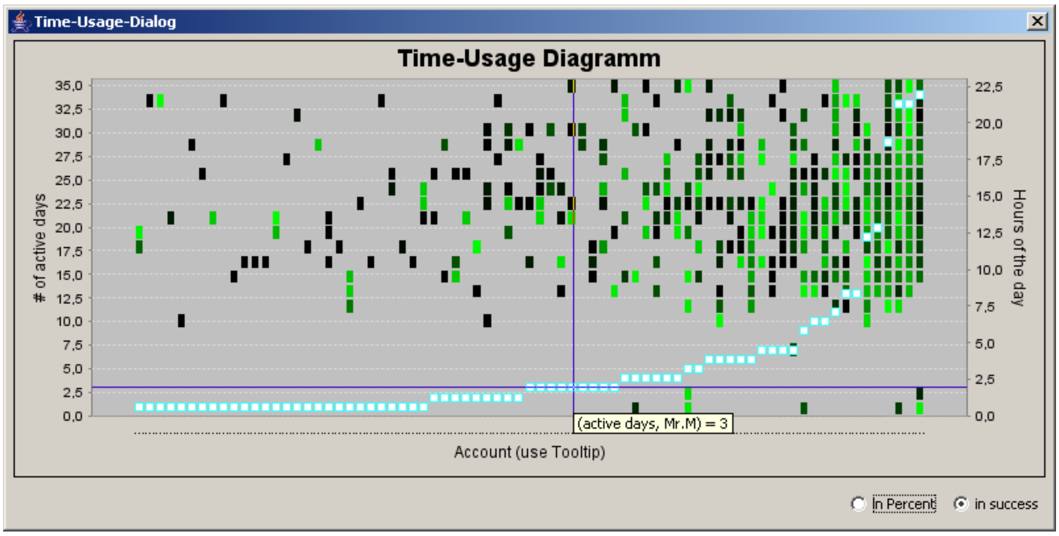
```
$E_v = \{e \in E | e \cap v \neq \emptyset\}$
$maxVol = 0$
$edgesFound = 0$
$maxVolEdge = \{\}$
$\forall e \in E_v$ :
  $if\ (volume(e) > \frac{volume(E_v)}{deg(v)} * \epsilon)\, then$
        $edgesFound = edgesFound + 1$
        $if\ (volume(e) > maxVol)$
            $maxVolEdge = e$
            $maxVol = volume(e)$
  $if\ (edgesFound = 1)\, then$
        $FP_v(v) = FP_v(v) + 1$
        $FP_e(maxVolEdge) = FP_e(maxVolEdge) + 1$
```

Algorithm 3: Prominent Edge Indicator

It may include two or up to three prominent edges. Imagine the case in which there is a promising share on the market that everyone expects to rise. Creating a second account, buying as much as possible of this share and selling it cheap to the primary account is a promising fraud strategy. Since this case is somewhat different from the Prominent Edge Indicator and has an even more specialised pattern, it will not be explained in detail in this overview.

3.5 Activity Indicator

Activity can be inspected from various angles. On the one hand, it can be interesting to compare the number of days in which an account was actively used (e.g. placing orders). It is very likely that the primary and secondary accounts will have taken contrary positions during this period. To give the user of the fraud detection system a general idea, the average or median can be given, too. Meanwhile an examination of the specific time of day at which an account was typically active may be interesting as well. This may help to detect rarely used accounts that have been trading during periods of low market activity. These account holders must either possess a different valuation of the shares (better information, novice user), restricted access times (working hours, etc.) or malicious intent to trade shares during times when no one else appears to want to engage in trading. In addition, simultaneous or consecutive use of two related accounts (i.e. primary and secondary) may warrant further investigation.



○ In Percent ● in success

Figure 1: Activity Indicator Example

We combined both variables (number of days the account was used as well as typical online times within the day) in one diagram. Figure 1 furnishes an example. The white squares indicate the number of days the account was used (depicted on the left scale axis). The vertical boxes are coloured to reflect the account holder's success rate (i.e. ratio of concluded transactions to offers placed) and arranged in rows, indicating the hour of the day when the offers were placed (depicted on the right scale axis). The accounts are too numerous to all be plotted on the x-axis; the cross can be used to obtain a tooltip for each square (see the lower middle area of Figure 1) that shows the actual value and the account. Again, an advantageous feature here is that the number of accounts can be reduced dramatically by filtering all accounts which are not adjacent to the vertex currently under investigation. This capability dramatically reduces complexity for the user.

The activity indicator correlates the activity times throughout the day per account with the days the account has actually been in use. The time events are derived from the events “transaction made” and “offer placed”.

Although it most likely is a very weak indicator, the time of registration is nonetheless worthy of consideration. Many beginners tend to overlook the registration timestamps for the two types of accounts. Primary accounts generally precede secondary accounts in the registration sequence.

4 Results

In this contribution, we provided a brief overview of the various possible indicators. Since the main idea of our work is to combine graph-based approaches with common transaction-based indicators, we will focus in this section on the Prominent Edge Indicator. Figure 2 shows a possible outcome of the anonymized market data visualized

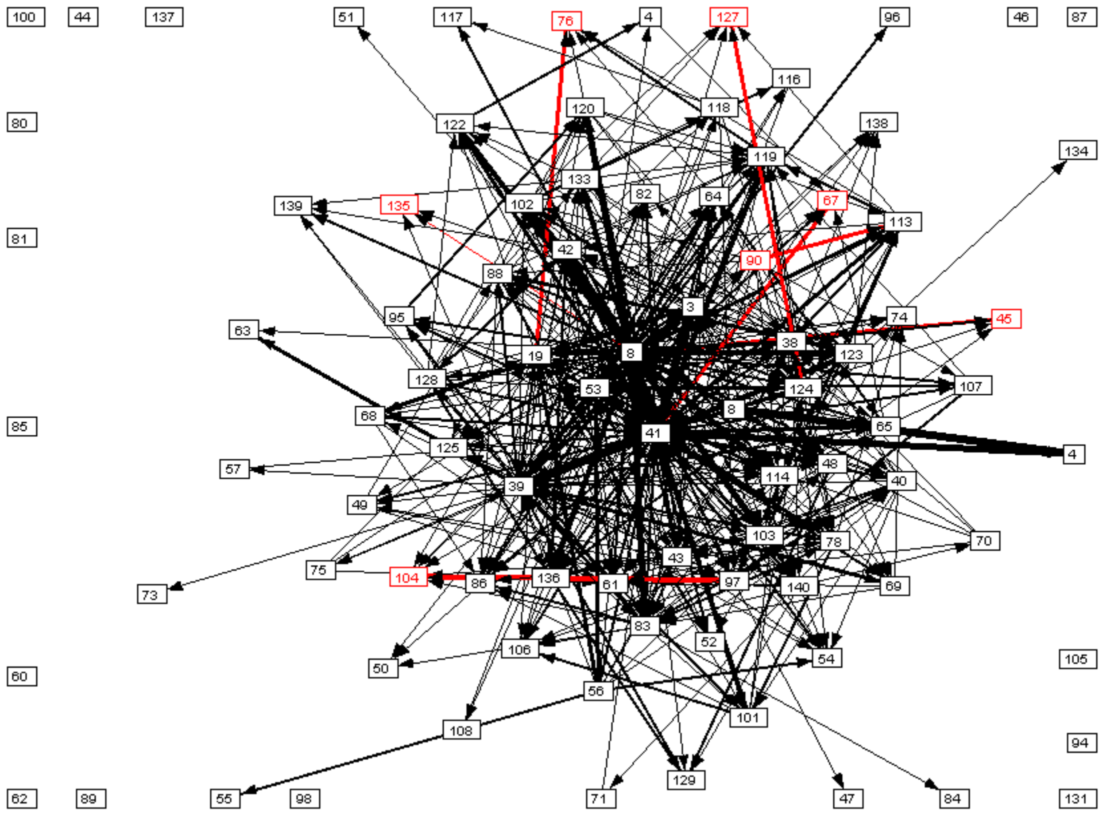


Figure 2: Prominent Edge Indicator

as a graph in force-directed layout (c.f. Fruchterman and Reingold, 1991). Every vertex in this graph represents an account, edges between accounts symbolise concluded transactions. The edges take the form of an arrow pointing from the seller to the buyer. The width of an edge is related to $\log \sum_i p_i * v_i$ (p price, v volume) over all transactions concluded between these two accounts. All suspicious vertices and edges detected by the Prominent Edge Indicator are highlighted in bold red typeface.

As expected, all suspicious vertices fall within the outer ring of the graph, since vertices are less connected here. Moreover, we can observe that most edges detected connect to vertices in slightly more embedded positions (inner rings) in the graph. These represent the corresponding primary accounts. From the three fraud cases known in the dataset, two were detected with this indicator. The third one does not display any prominent edges and thus cannot be detected with this indicator.

Regarding the other detected cases, it remains to be seen whether they are fraud-related or not. However, the probability for this increases if these vertices or edges are detected by other indicators as well.

5 Conclusion & Outlook

In this article, we showed that looking at a market as a graph may help to find correlations indicative of fraud and reduce runtime of the often complex search for interdependencies. The activity-, order book- and circular trading indicators are still works in progress. However, the rating approach facilitates the incorporation of additional indicators.

Alas, there is no perfect method for detecting fraud. In an individual wishes to commit fraud and knows how the supervisory entity searches for fraud, he can always circumvent detection via automated means. However, other market participants may observe suspicious events or correlations. If these are reported, even fine-tuned fraud can be exposed. Manual control will therefore always be important, visualisations combined with intelligent indicators may help to improve efficiency and accuracy. This example serves to show the combined strength of visualisation and network analysis.

Future work will include the incorporation of a machine learning algorithms such as genetic algorithms or reinforcement learning (Waltz and Fu, 1965; Barto et al., 1981) in order to fine tune the parameter set for the indicators, include more specialised fraud patterns and apply the broader set of indicators to other datasets. A promising method in this field is an agent-based market simulation in which only selected agents are equipped with fraud strategies. In such a dataset the exact number of fraud cases is known, which allows a better training of the learning algorithms.

References

Baquia (2002): "eBay estrena una nueva arma contra el fraude," online, `http://www.baquia.com/noticias.php?idnoticia=00008.20020607`.

Barto, A., R. S. Sutton, and P. S. Brouwer (1981): "Associative search network: A reinforcement learning associative memory," *IEEE Transactions on Systems, Man, and Cybernetics*, 40, pp. 201–211.

Brandes, U. and T. Erlebach (eds.) (2005): *Network Analysis - Methodological Foundations*, vol. 3418, Springer, tutorial.

BRD (1998): *Wertpapierhandelsgesetz*, Bundesrepublik Deutschland.

BRD (2002): "Börsengesetz," Bundesgesetzblatt, §4.

Forsythe, R., F. Nelson, G. R. Neumann, and J. Wright (1992): "Anatomy of an experimental Political Stock Market," *The American Economic Review*, 82(5), pp. 1142–61, `http://ideas.repec.org/a/aea/aecrev/v82y1992i5p1142-61.html`.

Fruchterman, T. M. J. and E. M. Reingold (1991): "Graph drawing by Force-directed Placement," *Software - Practice and Experience*, 21(11), pp. 1129–1164.

Jennifer Chasin, c. (2000): "BidPay.com's Fraud Detection System Uncovers International Online Auction Fraud Ring," online, `http://www.cardinalcommerce.com/articles/september\_2000/BidPay.com\%20Uncovers\%20Fraud\%20Ring.htm`.

Jungnickel, D. (2005): *Graphs, Networks and Algorithms*, *Algorithms and Computation in Mathematics*, vol. 5, 2nd edition, Springer, Heidelberg.

Library, F. (2003): "Internet Fraud related resources," online, `http://fbilibrary.fbiacademy.edu/Templates/B=internetfraud.htm`.

Luckner, S., F. Kratzer, and C. Weinhardt (2005): "STOCCER - A Forecasting Market for the FIFA World Cup 2006," in: *Proceedings of the 4th Workshop on e-Business (WeB 2005), Las Vegas, USA*.

Nelson, F., J. Berg, and T. Rietz (2003): "Accuracy and Forecast Standard Error of Prediction Markets," working paper, University of Iowa.

Ortner, G. (1996): *Experimentelle Aktienmärkte als Prognoseinstrument - Qualitätskriterien der Informationsverarbeitung in Börsen am Beispiel Political Stock Markets*, Ph.D. thesis, Universität Wien.

Philips, L. (2000): "The double metaphone search algorithm," *C/C++ Users J.*, 18(6), pp. 38–43.

Robertson Barrett, C. W. (2003): "Going, Going, Gone - Consumers Need to Take Precautions When Using Online Auctions," online, `http://www.consumerwebwatch.org/dynamic/e-commerce-investigation-going-going-gone.cfm`.

Russel, R. (1918): "INDEX (Soundex Patent)," U.S. Patent No. 1.261,167, 1-4.

Spann, M. and B. Skiera (2003): "Internet-based Virtual Stock Markets for Business Forecast," *Management Science*, 49(10), pp. 1310–1326, `http://www.extenza-eps.com/INF/doi/abs/10.1287/mnsc.49.10.1310.17314`.

Spann, M. and B. Skiera (2004): "Einsatzmöglichkeiten virtueller Börsen in der Marktforschung," *ZfB - Zeitschrift für Betriebswirtschaft (Ergänzungsheft)*, 74, pp. 25–48.

Wahab, M. S. (2004): "E-Commerce and Internet Auction Fraud: The E-Bay Community Model," working paper, Computer Crime Research Center, `http://www.crime-research.org/articles/Wahab1`.

Waltz, M. D. and K. S. Fu (1965): "A heuristic approach to reinforcment learning control systems," *IEEE Transactions on Automatic Control*, 10, pp. 390–398.

Modeling and Simulating Competition among e-Auction Marketplaces

Xin Chen[1], Christof Weinhardt[1], and Siegfried Berninghaus[2]

[1] Institute for Information Systems and Management,
Universität Karlsruhe (TH)
{christof.weinhardt,xin.chen}@iw.uni-karlsruhe.de

[2] Institute for Economic Theory and Operations Research,
Universität Karlsruhe (TH)
berninghaus@wiwi.uni-karlsruhe.de

Summary. Electronic auctions (e-auctions) are undoubtedly becoming an important part of e-commerce. In recent years, many similiar electronic auction marketplaces have been built. Sellers and buyers who want to trade via online auctions therefore have ample opportunity to compare and select marketplaces on which to trade. Marketplace operators thus compete with each other to attract participants. This paper models this competition among marketplaces and investigates two kinds of agent strategies: fixed strategy and adaptive strategy. An agent-based approach is applied to simulate the evolution process of the market structure based on the proposed model. Three conclusions can be drawn from the simulation results. Firstly, there exist equilibria in which marketplaces coexist and compete for participants. Secondly, such equilibria need not be unique. And thirdly, the model converges faster to equilibria if agents use the adaptive rather than the fixed strategy.

1 Competition among e-Auction Markets

As an important part of e-commerce, e-auction marketplaces have been drawing more and more attention. The fact that competition exists among buyers and sellers as well as among marketplaces that provide similar services for similar or even identical products is equally worthy of study. Marketplaces compete with each other for participants since both sellers and buyers have opportunities to compare and select a marketplace on which to trade. Intuitively, a seller prefers a marketplace with as many buyers and as few sellers as possible. Not surprisingly, a buyer prefers the inverse of these proportions. A participant can therefore be said to choose a market considering the choices of other participants. Conversely, his choice is also taken into account by other participants. Since each participant may join or leave a market at his own discretion, the market structure, as an aggregation of choices by all participants, is in a state of constant flux.

Empirically, we observe various market structure patterns under such competition in different countries. In the United States, for example, several marketplaces coexist

(Krishnamurthy, 2003). On the other hand, in Japan, Yahoo! Auction corners a market share of 95% and is the de facto monopolist. This market structure forced eBay to dissolve its business in Japan in March, 2002.[1] In mainland China, eBay, with over 80Interestingly, rather than driving its rivals out of the market, eBay's leading position is now being challenged by the rapidly growing new marketplace of Taobao.[2] It remains to be seen how China's market structure will evolve. It may develop such that one marketplace emerges as the sole winner. Alternatively, it could evolve into a duopolistic state in which two marketplaces (whose participants have no incentive to change from one to the other) coexist.

Although competition among markets has been a classical research topic in economics (see, for example, Gehrig 1998; Rochet and Tirole 2003; Kam et al. 2003), electronic markets, as an alternative to traditional markets, call for a better understanding of competition in virtual trading environments. There have been studies of competition among various Internet search engines (Mukhopadhyay et al., 2004), or among various intermediate service providers (Caillaud and Jullien, 2003), etc.

Concerning e-auction markets, some researches focus on designing better bidding strategies for buyers (see, for example, Milgrom and Weber 1982; Goeree and Offerman 2003) while others concentrate on mechanisms for sellers to attract more buyers (see, for example, Standifird 2002; Schoder and Haenlein 2004; Bandyopadhyay et al. 2005). Meanwhile, some papers investigate from the perspective of marketplace operators. However, those papers mostly analyse the differences in competition between traditional and electronic markets (Sivakumar, 2000), or zero in on the aspects of risk (Saeed and Leitch, 2003). The competition among the e-marketplaces themselves is seldom studied. Our work is most closely related to a recent theoretical model on competing online auction markets (Ellison et al., 2004). Ellison et al. conclude that in a setting in which the participants' decisions are only influenced by network effects, equilibria exist in which several operators coexist.

2 The Model

In this section, the competition among marketplaces is modelled using an agent-based approach. Each participant in a market is considered an independent agent. An agent should act presenting his corresponding participant's preferences and decisions (Kreps, 1990). Furthermore, each agent should also be an autonomous, computational entity that perceives its environment and acts without the intervention of humans or other systems (Weiss, 1999).

2.1 The Market Environment

Generally, a scenario of competition comprises a set of z markets that compete with each other. The markets are assumed to be identical in design. There are two types

[1] See `http://investor.ebay.com/ReleaseDetail.cfm?ReleaseID=77835`

[2] See `www.taobao.com`

of agents: sellers and buyers. Altogether there are n agents allocated in the markets, in which s agents are seller and b agents are buyers. Note that n, s, b are all constant numbers. The number of seller and buyer agents in a market M_i is denoted as $s(M_i)$ and $b(M_i)$ respectively. The distributed allocation of the agents in z markets can then be formalised as Equation 1.

$$n = s + b = \sum_{i=1}^{z} s(M_i) + \sum_{i=1}^{z} b(M_i) \tag{1}$$

We start with a simple scenario where $z = 2$, i.e. only two markets, M_1 and M_2, compete with each other. Figure 1 depicts the distribution of the agents. In each market, one auction takes place, and both auctions run in parallel. Each auction may contain multiple sellers and buyers. The mechanism of both auctions is a single-sided uniform price auction. It is a multi-unit auction in which units offered by the sellers are sold at a "market-clearing price" such that the total demand equals the total supply (Krishna, 2002). We adopt this mechanism for several reasons. Firstly, similar mechanisms are often applied, such as the so-called "dutch auction" in eBay[3] and the "call auctions" used by several securities exchanges[4]. Secondly, when buyers have unit demands, bidding with their true valuations is the dominant strategy. Here the market-clearing price is specified to be the highest rejected bid. The number of sellers satisfies $s < \lfloor n/2 \rfloor$, so that the price determination is ensured to be executable in at least one of the two marketplaces.

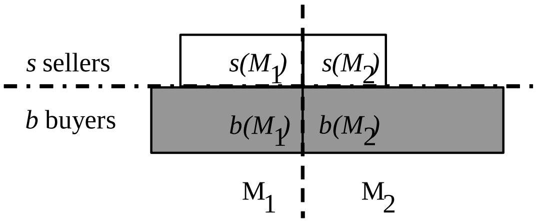


Figure 1: Distributed allocation of agents in two competing markets

2.2 The Agents

In each round, each buyer has unit demand and each seller has unit supply of item. All items demanded or offered are identical. Cross bidding or selling at two markets is not allowed. Each buyer has a private valuation on the item in demand, and the buyers' valuations are independently drawn from a uniform distribution on support $[0, \bar{c}]$, $\bar{c} > 0$. The sellers are assumed to be price-takers, i.e. they all have a reserve price of zero. Thus, the buyers have no incentive to compare the offers and improve their own utilities by trading with a different seller. According to their own strategies,

[3] See `http://pages.ebay.com/help/buy/buyer-multiple.html`

[4] Both New York Stock Exchange and Frankfurt Stock Exchange use opening call auctions.

agents decide whether to stay in the current market or move to the other market. Thus, an agent's decision space D contains only two elements, i.e. $D = \{stay, move\}$. Each agent makes decisions privately and independently.

Agents are assumed to be "boundedly rational" in terms of decision-making (Simon, 1955). Under this concept, agents evaluate the markets and make decisions in accord with the following hypotheses. Firstly, agents are myopic (myopia hypothesis). In such a repeated game with a large population, it is computationally difficult for an agent to compute the best strategy in anticipation of the other agents' possible actions. Therefore, agents are assumed to be myopic, meaning that there is no memory of information from the previous rounds, and no speculation for the next rounds. Any behaviour on the part of agents is based solely on the knowledge acquired in the current round.

Secondly, agents are inert (inertia hypothesis). When the evaluation of the two markets comes out equal, i.e. the agent is indifferent towards the markets, inertia causes the agent to stay at the current market.

Last but not least, the mutation hypothesis plays a role in agent behaviour. On one hand, agents are limited by constraints in information, knowledge and cognitive capacity, etc. On the other hand, barriers prevent agents from changing marketplaces. Obstacles may include transaction costs, transfer costs, learning costs associated with obtaining familiarity with another marketplace platform, etc. Therefore, there is a small probability of agents' behaviour deviating from their decisions in the rational case.

2.3 The Auction Process

The auction process is a repeated execution of auctions organised by rounds. In each round, one auction in each market takes place via the execution of three sequential steps, as shown in Fig. 2. One additional step is necessary for initialisation. The two auctions are synchronised from step to step. The two markets have the same auction process.

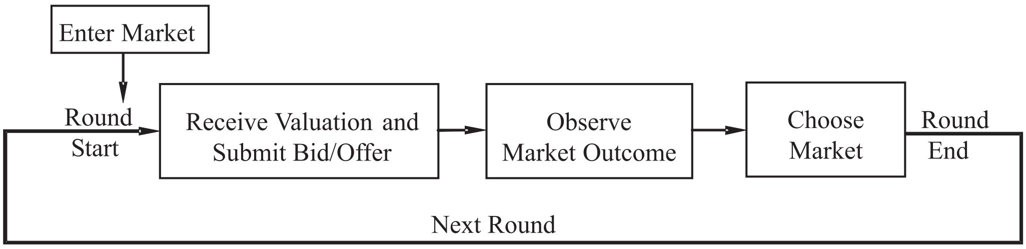


Figure 2: The auction process

Step 0: Enter Market

In the beginning, agents radomly enter one of the two markets.

Step 1: Receive Valuation and Submit Bid

Each buyer receives a private valuation from a given distribution and bids with this valuation. Each buyer submits one and only one bid in a round.

Step 2: Observe Market Outcome

After all buyers in a market have submitted their bids, the selling price and the winners are determined. The items are then allocated. The price as well as the current number of sellers and buyers in both markets are published as the "market outcome" and are observed by each agent.

Step 3: Choose Market

Based on the market outcome, each agent independently decides whether to stay in the current market or to join the other market in the next round. Each agent may make this decision only once per round.

3 Decision Criteria and Agent Strategy

According to the auction process, in each round, agents must choose a marketplace before receiving their private valuations. One explanation for this is that buyers need to join a marketplace and inspect the items on offer to learn their evaluations. This is especially applicable when the items offered by sellers in different rounds are different. Under such conditions, agents can not evaluate the expected payoff in a market by calculating with their valuations. In a closely related model by Ellison et al., in which the auction mechanism and the distribution of agent valuations are the same, it has been shown that an agent's expected payoff in a market can be transformed to be presented as a proportion of the number of sellers to the number of buyers in that market (hereafter referred to as seller-buyer ratio) (Ellison et al., 2004). This is easy to understand. If the number of buyers is fixed in a given market, the price will increase if the number of sellers decreases. If the number of sellers is fixed, the price might also increase if the number of buyers increases. In this paper, we use the seller-buyer ratio as the basis of decision-making and agent strategies. $\mathit{\Gamma}(M_i) = s(M_i)/b(M_i)$ denotes the seller-buyer ratio of market M_i. Since agents intuitively prefer to trade in a market which provides a higher expected payoff, a buyer prefers a market in which the $\mathit{\Gamma}$ is as large as possible. On the contrary, a seller prefers a market with a smaller $\mathit{\Gamma}$.

For any agent a_k, the market in which a_k is currently is denoted as located as local market M_x, and the other market is denoted as his remote market M_y. Thus, the difference in ratio from agent a_k's point of view is the seller-buyer ratio of his local market obtained by subtracting the remote market seller-buyer ratio from the local market counterpart, as shown in formula (2).

$$\mathit{\Delta\Gamma}(a_k) = \mathit{\Gamma}(M_x)\,|_{a_k \in M_x} - \mathit{\Gamma}(M_y)\,\Big|_{a_k \notin M_y} \tag{2}$$

Based on the seller-buyer ratio, sellers and buyers appear to make decisions differently. Formulas (3) and (4) show the decision-making strategies of agent a_k as a seller (denoted as a_k^S) and as a buyer (denoted as a_k^B), respectively. A rational seller chooses to stay in the current market if the ratio difference is not positive, i.e. the local market's ratio is less than or equal to the remote market's ratio; otherwise, the seller decides to move to the remote market in the next round. A rational buyer behaves in the opposite way.

$$a_k^S = \begin{cases} stay, & \Delta\Gamma(a_k^S) \leq 0 \\ move, & \Delta\Gamma(a_k^S) > 0 \end{cases} \tag{3}$$

$$a_k^B = \begin{cases} stay, & \Delta\Gamma(a_k^B) \geq 0 \\ move, & \Delta\Gamma(a_k^B) < 0 \end{cases} \tag{4}$$

Formulas (3) and (4) assume that all agents are rational. According to the mutation hypothesis, agents are boundedly rational and may change their strategies at random. We refer to an agent's rational decision as an intention and suppose that a mutation of decision occurs only when an agent's intention is to move. How rational an agent is can be formalised as a parameter of mutation rate p_m, which stands for the probability of an agent's decision being consistent with his intension; $p_m \in [0, 1]$. $p_m = 0$ represents the situation in which each agent always sticks to his local market while $p_m = 1$ indicates that all agents behave purely according to their rational decisions.

In this paper, we simulate agents' behaviour according to two strategies: fixed and adaptive. Agents using a fixed strategy act with a mutation rate of fixed value. In contrast, if an agent uses an adaptive strategy, the mutation rate p_m varies according to the changes of the ratio difference $\Delta\Gamma$, as shown in formula (5).

$$p_m = \mu * |\Gamma(M_1) - \Gamma(M_2)| \tag{5}$$

p_m is not static; it reflects the rationality level of agents dynamically. This is actually intuitive: a significant ratio difference indicates a possible significant payoff difference with the result that agents are more willing to move. μ is a scaling coefficient in the field of $[0, 1]$, indicating the sensitivity of variance of bounded rationality.

4 The Simulation Results

We use the *meet2trade* tool suite as the supporting platform for implementing the simulation based on the model proposed in Section 3.[5] meet2trade is a generic trading platform that facilitates the designing, building and testing of electronic markets, ranging from non-formalised bilateral negotiations to formalised auction mechanisms (Weinhardt et al., 2005). The design of the simulation framework and the implementation can be seen in Chen et al. 2005.

Two series of simulations were conducted. Series I was conducted to observe the evolution process of the market structure and see if equilibrium in the form of a market duopoly occurs (in which the two markets coexist and the market structure is stable). This series consists of two different settings, as listed in Table 1. Each setting designates

[5] meet2trade is a generic trading tool suite developed by the Chair of Information Management and Systems, Universität Karlsruhe (TH), Germany. See `http://www.meet2trade.com`

the agents in their initial locations in the two markets. Besides, agent strategies and related mutation rates are also defined. Each setting is simulated for 20 runs. In each run, simulation may end at one of three states: a monopolistic state, in which only one market survives; a state of equilibrium, in which a stable duopolistic market structure is observed; or an unstable state, in which up to 40 simulation rounds have been run and neither of the above two states can be observed.

Table 1: Simulation settings for Series I

Setting	n	$s(M_1)$	$b(M_1)$	$s(M_2)$	$b(M_2)$	Strategy	p_m	Runs	Max. Rounds
1	40	3	8	13	16	fixed	0.111	20	40
2	600	40	120	80	360	fixed	0.111	20	40

Figure 3 shows the state of the market structure when the simulations terminate. Under Setting 1, 9 out of 20 runs converge to equilibria in the form of a duopoly, while 4 runs under Setting 2 also converge to equilibria. All together, 32.5% of the simulation runs converge to equilibria, which indicates that equilibrium occurs rather frequently. Therefore, the conclusion can be drawn that in two competing markets where agents make decisions based upon these markets' seller-buyer ratios, there exists equilibrium in the form of market duopoly, in which no agent is motivated to leave his current market and join the other one.

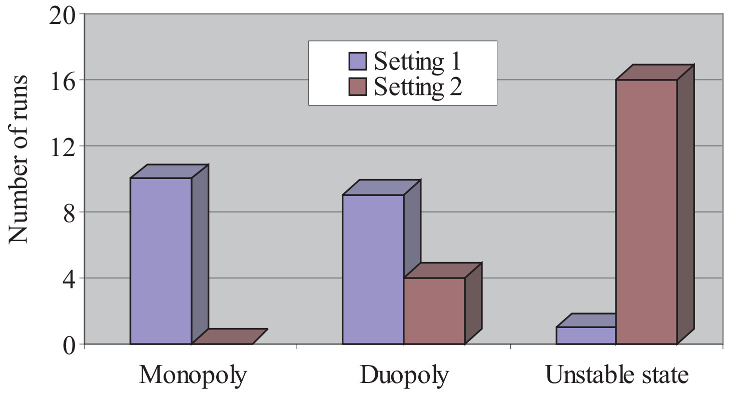
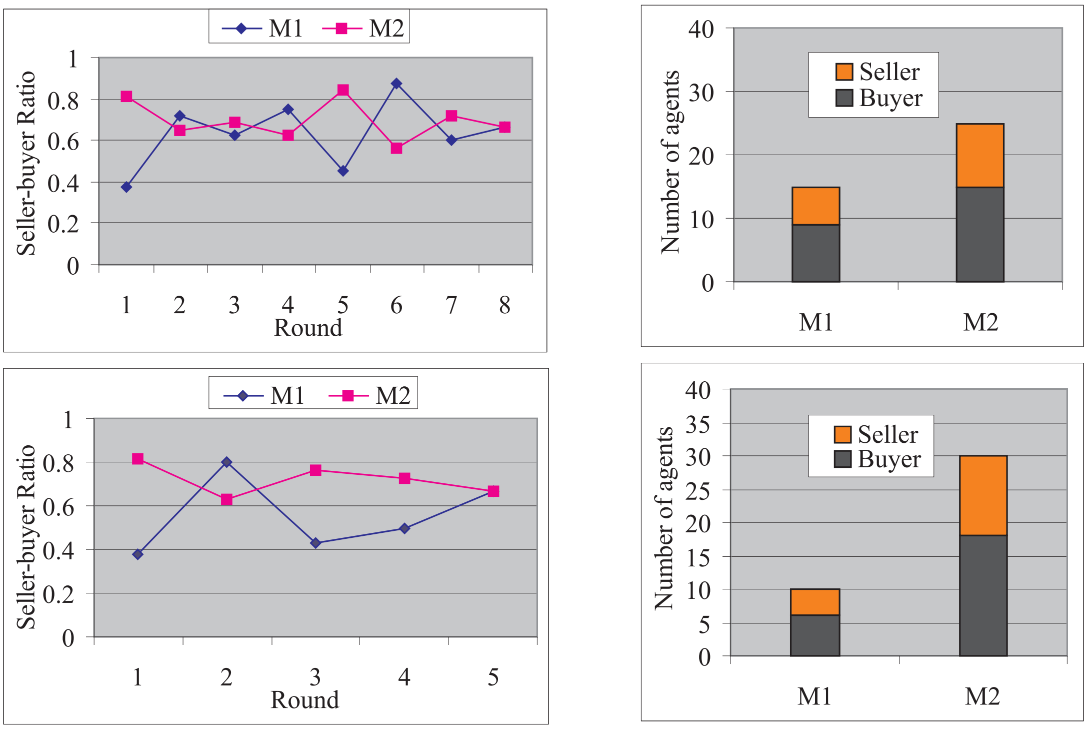


Figure 3: End states of simulations

Our next step is to investigate the convergence process towards equilibrium and the market structure at equilibrium. We look at the simulation runs that achieved equilibrium in the form of duopoly under Setting 1. Fig. 4(a) shows two of those runs and depicts the dynamics of the seller-buyer ratios in the two markets. Under the same setting (Setting 1), in Round 1 of the two runs, $\Gamma(M_1) = 0.813$ and $\Gamma(M_2) = 0.375$. By the end of the simulations, the ratios converge to an identical number, i.e. $\Gamma(M_1) = \Gamma(M_2) = 0.667$. However, Fig. 4(a) shows that the converging paths to achieve equilibrium and the number of rounds needed for such convergence are different. The simulation run in the upper row converge to equilibrium after 8 rounds, while the simulation run in the bottom row converge within 5 rounds. The ratios of the two markets during the convergence are also obviously different. Furthermore, Fig. 4(b) manifests the stable market structures at equilibria in those two simulation runs. In the simulation run in the upper row, M_1 contains 6 sellers and 9 buyers, and M_2 contains 10 sellers and 15 buyers. In the simulation run below, the market structure differs significantly with only 4 sellers and 6 buyers in M_1, while M_2 is much larger, with 12 sellers and 18 buyers.



(a) Different converging paths towards equilibria

(b) Different market structures at equilibria

Figure 4: Existence of equilibria on the market structure

We have also studied the converging process and market structure in simulations conducted at Setting 2. Among larger populations of agents, similar results can be observed. This leads to the second conclusion that equilibrium in the form of duopoly may not be unique. There may in fact be multiple equilibria, in which market structures differ from one another.

The second series was designed to facilitate the observation of how the market structure evolves when the agents are not purely rational. This series differs from the first series in that agents use the adaptive rather than the fixed strategy. This series contains five settings (see Table 2), each setting has 600 agents and the population is stationary, as in Setting 2. From Setting 3(a) to 3(e), the proportion of the total number of sellers to buyers is kept to $1/4$, with $s = 120$ and $b = 480$. According to the empirical data collected from eBay (Dietrich and Seese, 2004), during a randomly chosen period of time, the seller-buyer ratio in the market falls in the field of [0.204, 0.503]. Therefore, the setting on agent numbers is appropriate.

Table 2: Simulation settings for Series II

Setting	n	$s(M_1)$	$b(M_1)$	$s(M_2)$	$b(M_2)$	Strategy	μ	p_m	Runs	Max. Rounds
3(a)	600	40	80	80	400	adaptive	0.5	dynamic	20	40
3(b)	600	40	280	80	200					
3(c)	600	40	240	80	240					
3(d)	600	40	200	80	280					
3(e)	600	40	120	80	360					

Setting 3(a) is first simulated and compared to simulation runs with Setting 2. Those two settings differ only in terms of agent strategies. The initial value of p_m under the adaptive fixed strategy is 0.111, which is equivalent to the p_m in the fixed case. Thus, simulations conducted under both strategies have the same ”starting point”.

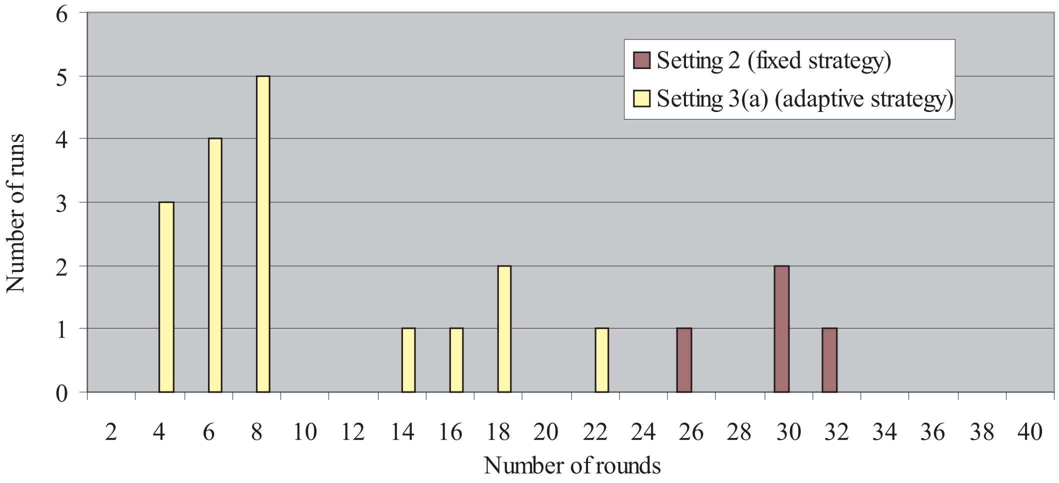


Figure 5: Simulation results under heterogeneous agent strategies

We observed in Fig. 5 that the simulation runs with adaptive agents have much higher probability of converging to equilibria than agents operating under stationary strategy. Under the fixed strategy, only 20% of the simulation runs converge to equilibrium, within an average of 29.5 rounds. In contrast, the adaptive strategy results in 90% of the simulation runs converging to equilibria within 24 rounds. Notably, 60% of the simulations converge within 8 rounds, which is much lower than the average number of rounds in the fixed case.

Additional simulations using the Settings 3(b) to 3(e) were conducted to examine whether the above results hold universally under heterogeneous initial distributions of agents. From Setting 3(b) to Setting 3(e), the initial seller-buyer ratio in M_1 increased while the ratio in M_2 is decreased. The ratio difference of the two markets peaked in Setting 3(d). Other parameters remainded the same. Figure 6 shows the number of rounds needed to achieve equilibrium under those different settings. In simulation runs with all four settings, equilibrium was achieved quite often. In particular, all the simulations runs under Settings 3(b), 3(c) and 3(d) converge to equilibrium. Moreover, on average, 87.5% runs converged to equilibrium within 10 rounds. This is consistent with the result yielded by Setting 3(a). Therefore, we can conclude evolves more easily and rapidly to equilibrium in the form of a duopoly when agents make decisions under the adaptive strategy rather than under the fixed strategy. This conclusion is independent of the initial states of the market structure (i.e. the initial distribution of agents) at the beginning of the simulations.

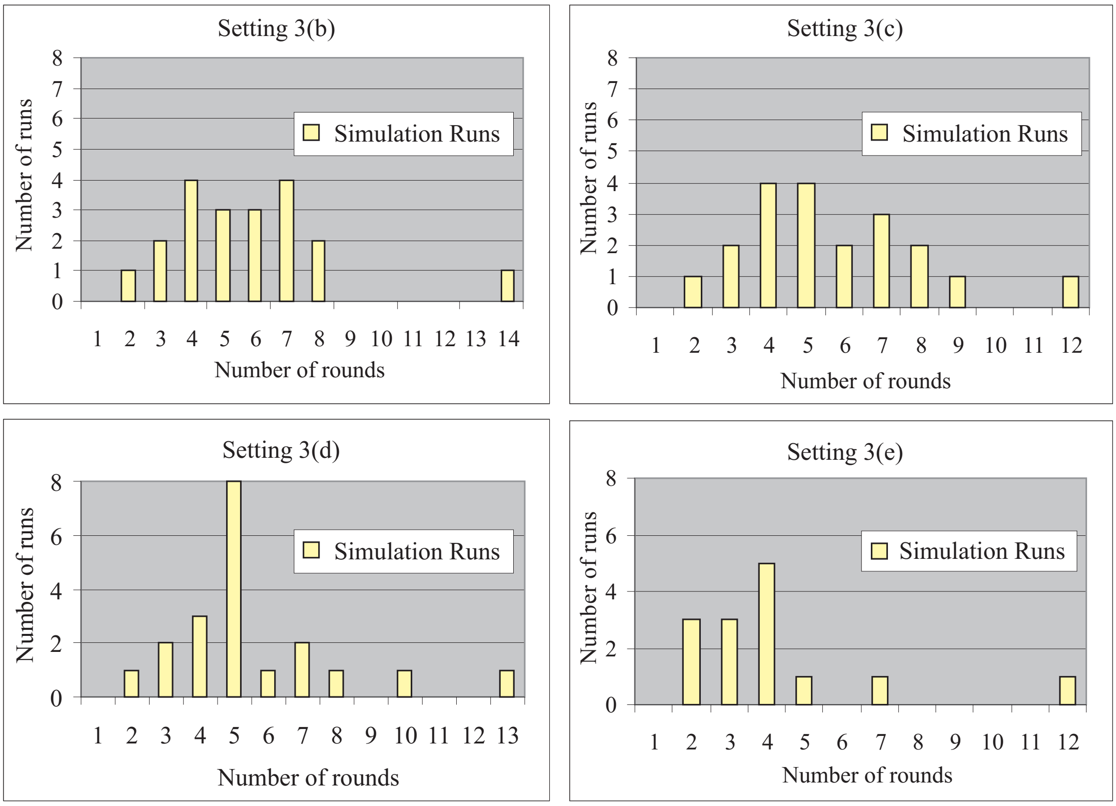


Figure 6: Simulation results under heterogeneous initial agent distributions

5 Conclusion and Outlook

In this paper, a model is proposed to investigate the competition among e-auction marketplaces. The simulation results proved that there equilibrium occurs in two co-existing competing markets, in which participating agents remain stationary. The simulations also illustrate the non-uniqueness of such equilibria. This result strongly supports the argument in Ellison and Fudenberg 2003 asserting that there is a broad plateau of equilibria in two competing markets. Furthermore, we demostrated that the market structure tends to evolve into equilibrium in the form of duopoly rather than monopoly when agents use the adaptive strategy.

Using the agent-based approach, our model can easily be extended to include more sophisticated agent strategies and competition scenarios, which is relatively difficult to achieve via theoretical modelling. For example, the model can be extended to simulate the competition among multiple markets. Moreover, the assumption that all institutions in the markets are identical can be also relaxed. Simulations can be conducted with marketplaces with different transaction cost policies. Concerning the individual agents, agent strategies can be further differentiated according to various risk-types. Simulations can be also conducted to observe how the market structure evolves with heterogeneous populations. These extensions can be further combined to model more complicated scenarios.

References

Bandyopadhyay, S., J. Barron, and A. Chaturvedi (2005): "Competition Among Sellers in Online Exchanges," *Information Systems Research*, 16(1), pp. 47–60.

Caillaud, B. and B. Jullien (2003): "Chicken & Egg: Competition among Intermediation Service Providers," *RAND Journal of Economics*, 34(2), p. 309.

Chen, X., J. Maekioe, and C. Weinhardt (2005): "Agent-based Simulation on Competition of e-Auction Marketplaces," in: *International Conference on Intelligent Agents, Web Technologies and Internet Commerce*.

Dietrich, T. and D. Seese (2004): "Der Handel bei eBay.de," 2004(1), pp. 17–30.

Ellison, G. and D. Fudenberg (2003): "Knife-Edge Or Plateau: When Do Market Models Tip?" *The Quarterly Journal of Economics*, 118(4), pp. 1249–1278.

Ellison, G., D. Fudenberg, and M. Moebius (2004): "Competing Auctions," *European Economic Review*, 2(1), pp. 30–66.

Gehrig, T. (1998): "Competing markets," *European Economic Review*, 42(2), pp. 277–310.

Goeree, J. K. and T. Offerman (2003): "Competitive Bidding in Auctions with Private and Common Values," *Economic Journal*, 113(489), pp. 598–613.

Kam, T.-K., V. Panchapagesan, and D. G. Weaver (2003): "Competition among markets: The repeal of Rule 390," *Journal of Banking and Finance*, 27(9), pp. 1711–1736.

Kreps, D. (1990): *A Course in Microeconomic Theory*, Princeton Univ. Press.

Krishna, V. (2002): *Auction Theory*, Academic Press.

Krishnamurthy, S. (2003): *A Comparative Analysis of eBay and Amazon*, Idea Group Publishing.

Milgrom, P. R. and R. J. Weber (1982): "A Theory of Auctions and Competitive Bidding," *Econometrica*, 50(5), pp. 1089–1122.

Mukhopadhyay, T., U. Rajan, and R. Telang (2004): "Competition between internet search engines," in: *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, IEEE, pp. 216–225.

Rochet, J.-C. and J. Tirole (2003): "Platform Competition in Two-Sided Markets," *Journal of the European Economic Association*, 1(4), pp. 990–1029.

Saeed, K. and R. Leitch (2003): "Controlling Sourcing Risk in Electronic Marketplaces," *Electronic Markets*, 13(2), pp. 163–173.

Schoder, D. and M. Haenlein (2004): "The relative importance of different trust constructs for sellers in the online world," *Electronic Markets*, 14(1), pp. 48–57.

Simon, H. (1955): "A Behavioral Model of Rational Choice," *The Quarterly Journal of Economics*, 69(1), pp. 99–118.

Sivakumar, V. (2000): "Competition across channels: do electronic markets complement or cannibalize traditional retailers," *Proceedings of the twenty first international conference on Information systems*, pp. 513–519.

Standifird, S. (2002): "Online Auctions and the Importance of Reputation Type," *Electronic Markets*, 12(1), pp. 58–62.

Weinhardt, C., C. van Dinter, K. Kolitz, J. Maekioe, and I. Weber (2005): "meet2trade: A generic electronic trading platform," in: *4th Workshop on e-Business(WEB 2005).*

Weiss, G. (1999): *Multiagent systems: a modern approach to distributed artificial intelligence*, MIT Press.

Communication and Control: Joint Treatment of Application-Specific Behavior and Communication Constraints in VANETs

Moritz Killat[1], Hannes Hartenstein[1], and Karl-Heinz Waldmann[2]

[1] Institute of Telematics,
Universität Karlsruhe (TH),
killat@tm.uni-karlsruhe.de, hartenstein@rz.uni-karlsruhe.de

[2] Institute for Theory of Economics and Operations Research,
Universität Karlsruhe (TH),
waldmann@wior.uni-karlsruhe.de

Summary. A layerist approach separates the design of a distributed application from the design of the communication system. The communication system provides generic communication services with usually sufficient quality for standard application classes. Recent developments in wireless networks, however, come up with applications whose constraints might exceed the physical capabilities of the provided services when not treated jointly. We look at *Vehicular Ad-Hoc Networks (VANETs)* that are assumed to be a fundamental next step in order to improve the safety of driving vehicles. Being supported by a so called *Driving Assistant System (DAS)* the driver gets assisted to take actions that might, for instance, improve the flow of traffic or that guide him or her in situations of danger. Such services are subject to the peculiarities of VANETs that especially suffer from adverse radio channel conditions. Thus, at least two sources of uncertainty exist: behavior of the driver and 'behavior' of the radio communication. To assess the quality of the communication system one eventually has to take into account how much the communication system helps to increase safety. Based on a recent work of Dolgov and Laberteaux we address this question with a Markov Decision Process (MDP) toy example that shows the influence of unreliable communications on the vehicle collision probability. Second, we show that the decision of re-broadcasting a messages might also be viewed as an MDP. The structure of the problems shows an analogy to *Market Engineering* where *agents* act within the frame of a given *market structure* in order to achieve the best *performance*, i.e., in our scenario the driver's incentive for increased safety.

1 Introduction

We have witnessed a fast evolution of automotive safety systems. It had started with mechanisms like seat belts or head-rests that passively protect the driver in situations of danger. Technological advances led to safety systems that are also actively involved as, for instance, the anti-lock braking system. Although vehicles got more and more equipped with further electronic developments all the collected data were only measured by the vehicle itself that, consequently, restricted their validity to the close surrounding. The latest trend in vehicle safety systems tries to overcome this barrier by exchanging informations between vehicles via wireless communication. By this means, electronic *agents* that are installed in the vehicles might be able to analyze the traffic situation in a way that exceeds the driver's capabilities. Furthermore, the agents might notify each other on certain incidents in order to defuse crucial situations in their origin. Based on these informations an agent might warn the driver of taking unsafe actions or assist him/her in dangerous situations.

Even if an agent has a clear picture of the current traffic situation a proper driver assistance becomes highly complicated by the unawareness of the behavior of other vehicles in the next seconds. What are the neighboring vehicles' reactions on notification of a certain incident? And how should an individual adjust its choice of reaction in order to consider the initial incident on the one hand but also a potential "follow-up" accident with the neighboring vehicles on the other hand? In *Vehicular Ad-Hoc Networks (VANETs)* this issue is additionally stressed as we have to deal with another source of uncertainty. Due to several reasons as, e.g., the high number of objects that are able to degrade the quality of the transmitted signal, scenarios in VANETs present adverse channel conditions that results in unreliable communication (Torrent-Moreno, Killat, and Hartenstein, 2005). Consequently, our concern is not only the choice of reaction taken by neighboring vehicles but also if the neighboring vehicles are aware at all of the just occurred incident, i.e., if the vehicles have received a specific message. This issue, again, has to influence the decision process: what would be the best reaction no matter if the vehicles in the vicinity have or have not received the notification, i.e., if they are about to change their current driving behavior.

Clearly, modeling and simulating accurate driver behavior in a realistic, i.e., large-scale, situation including an accurate modeling of the radio channel and the communication system represents a complex or even impossible task. On the other hand, the communication system has to be assessed by the application-based metric of how much the system increases safety (or comfort). As a starting point for such an application-based evaluation of the communication system we look at a toy example expressed as an Markov Decision Process that involves the two types of uncertainty: the uncertainty about the behavior of other vehicles (or their drivers) and the uncertainty about the reliability of the communication.

As a second approach we address another decision problem: based on context information known to the application and/or the communication system, when should a VANET node re-broadcast a message to achieve a 'pseudo-reliability' of reception of the message for all neighboring cars within a certain time limit?

In order to address these issues we need to take a 'super-position' that allows a simultaneous consideration of the application-specific behavior and the constraints of the communication medium. The structure of the problem description sketched so far shows an analogy to *Market Engineering* and, hence, suggests an approach in this context. Indeed, we can identify the driver's incentive to participate in such a system in order to increase his/her quality of driving, for instance, in terms of safety.

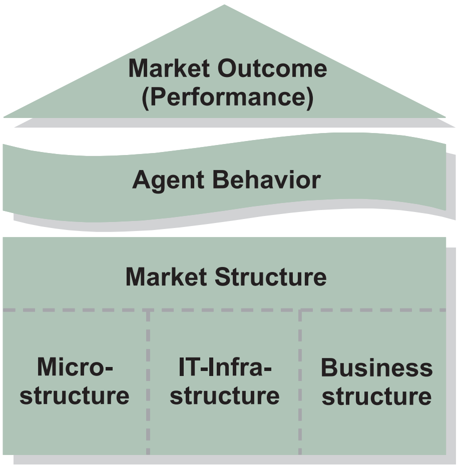


Figure 1: Structure of *Market Engineering*

Furthermore, the conceptional outline of Market Engineering, as depicted in Figure 1, shows various similarities to what we have discussed above: *Agents* act upon a *Market Structure* in order to achieve a good *Performance*. Likewise, Driving Assistant Systems operate within the system 'traffic' in order to improve the quality, respectively, the safety of driving vehicles. The Market Structure, again, consists of three pillars whereas the third, the *Business Structure*, is not well defined but leaves scope for development. Compared with this, the *IT-Infrastructure* essentially facilitates the agents' behavior in the sense of interaction. Their autonomy is solely restricted by the market defining rules kept in the first pillar, the *Microstructure*. In the vehicular scenario, however, the microstructure comprises considerations in order to overcome the unreliable communication.

This paper explains the difficulty of keeping a strict separation of communication and application layer by means of an example in Vehicular Ad-hoc Networks. We point out that mutual influences of the application and communication system forces us to take a kind of 'super-position' that simultaneously allows us to consider the constraints and to meet the requirements. On basis of a previous work of Dolgov and Laberteaux (Dolgov and Laberteaux, 2005), we demonstrate in Section 2 the effect of an unreliable communication system to the addressed application. In Section 3 we tackle the problem on behalf of the application that demands the communication system to provide certain services. Finally, we conclude this paper in Section 4.

2 Impact of Unreliable Communication

In dangerous situations in the daily road traffic, a driver might have several options to react on a certain incident but only a few might really diminish the probability of a following collision. Due to the multitude of vehicles in the vicinity and due to all their possibilities of reaction, the problem of choosing an appropriate reaction is very complex. In this section we substantiate this statement by referring to a drastically simplified model that has been proposed in (Dolgov and Laberteaux, 2005). By means of *Markov Decision Processes (MDPs)* we present and solve the mathematical problem that let us take the expected best reaction and prove the suitability using a simulation. Finally, we discuss the adverse conditions in VANETs, expressed in an unreliable communication, that significantly affects the probability of collisions.

2.1 Problem statement

In the following, we digress from a vehicular scenario and take a look at a square consisting of n fields on which two players move. While one of the players is uncontrolled and does random movements to the North, East, South and West we will assign a strategy to the second player in order to prevent collisions between both of them. Both players must move simultaneously and are not allowed to cross the borders of the square. As both players have n possibilities of placement, the entire system can be described in n^2 system states. If there would exist a partial order over the set of states in the sense of decreasing values for states with increasing probability of collision (lowest value for a collision), the controlled player could simply choose an action leading to a most valuable following state. According to Markov Decision Processes by means of which we address this issue, we derive the value of a state from the maximum reward that a player could obtain in the current state and from the expected total discounted reward of following states (for a more detailed introduction on Markov Decision Processes we refer to Appendix A). In mathematical terms, the value v_i of a state i is (usually) given as the unique solution of the following functional equation (so-called optimality equation)

$$v_i = \max_{a \in A} \{r_{ia} + \gamma \sum_{j \in S} p_{iaj} v_j\}, \tag{1}$$

where A denotes the set of possible actions and S the set of states. The one-stage reward for taking action a in state i is expressed by r_{ia}, and p_{iaj} describes the transition probability of going to state j when taking action a in state i.

A map $f : S \to A$ which specifies the action $a \in A$ to be taken in state $i \in S$ is called a decision rule. A (Markov) policy $\pi = (f_0, f_1, \ldots)$ is then a sequence of decision rules specifying the action $a_n = f(i_n)$ to be taken in state i_n at time n. Mainly, however, one is interested in stationary policies $\pi = (f, f, \ldots)$, for which we also write f.

It is well known in dynamic programming, that each decision rule formed by actions each maximizing the right hand side of (1) is optimal, i.e. leads to v_i for all initial values $i \in S$.

The optimality equation (1) is usually solved by value iteration (combined with an extrapolation), by policy iteration, or, as in our case, by linear programming.

2.2 Simulation

Our simulation underlies a 5×7-square, i.e., $n = 35$ fields, that yields $n^2 = 1225$ system states. The transition probabilities between system states depend on the type of field on which the uncontrolled player is located at the moment of decision making. As the following field for the controlled player, and so one part of the next system state, is determined by the chosen action a, the following system state only depends on the random movement of the uncontrolled player. Therefore we have to distinguish three cases for all $i, j \in S$ and $a \in A$:

$$p_{iaj} = \begin{cases} 0.5 & i \text{ comprises a corner field for the uncontrolled player} \\ 0.33 & i \text{ comprises an edge field for the uncontrolled player} \\ 0.25 & \text{otherwise} \end{cases}$$

The reward function is kept quite simple as we basically solely pay attention to collisions in terms of a negative reward. A collision not only occurs if the current system state implies both players to stay on the same field but also if they have changed their fields in a single moving step (as they would collide in between). In contrast to the former, the latter case can be anticipated before an actual accident happens and therefore demands a negative sanction of an action that might lead into such a collision. Naturally, this reward needs to be less harmful as an accident has not taken place so far. Thus, we penalize such a possible collision attenuated by the probability q with which the uncontrolled player would move to the current field of the controlled player. Finally, this modeling needs to be extended by a third, most restrictive punishment as we have to prevent illegal system actions taken by a player. Exemplarily, we want to suppress a movement to the north if the player is already placed at the upper edge of the square (he/she would illegally leave the square). Hence, we place a huge burden on a player who is about to choose an illicit action. Summing up, we use the following reward function

$$r_{ia} = \begin{cases} R_{collided} & \text{system state } i \text{ implies both player are located on the} \\ & \text{same field} \\ q \cdot R_{collided} & \text{action } a \text{ in system state } i \text{ leads the player to the current field} \\ & \text{of the other player} \\ R_{illegal} & \text{action } a \text{ is not allowed in system state } i \\ 0 & \text{otherwise} \end{cases}$$

with $R_{illegal} \ll R_{collided} < 0$ and q denotes the moving probability of the uncontrolled player.

Finally, we assume the starting state of the system to be uniformly distributed, i.e., $\alpha_i = \frac{1}{n^2} = \frac{1}{1225}$, the set of possible actions to comprise movements to the NORTH, EAST, SOUTH and WEST and the discount factor to be set to $\gamma = 0.9$ (cp. Appendix A).

Having applied these settings to the optimality equation (see Equation 1), we derive a policy by choosing that action a that leads us to the expected most valuable following system state, i.e., being in system state i, $a = \arg\max_a [r_{ia} + \gamma \sum_j p_{iaj} v_j]$. Exemplarily, a visualization of the computed policy for one possible placement of the uncontrolled player is illustrated in Figure 2(a). Furthermore, we examined the benefit of the computed strategy by means of a simple simulation in which both players made 10,000,000

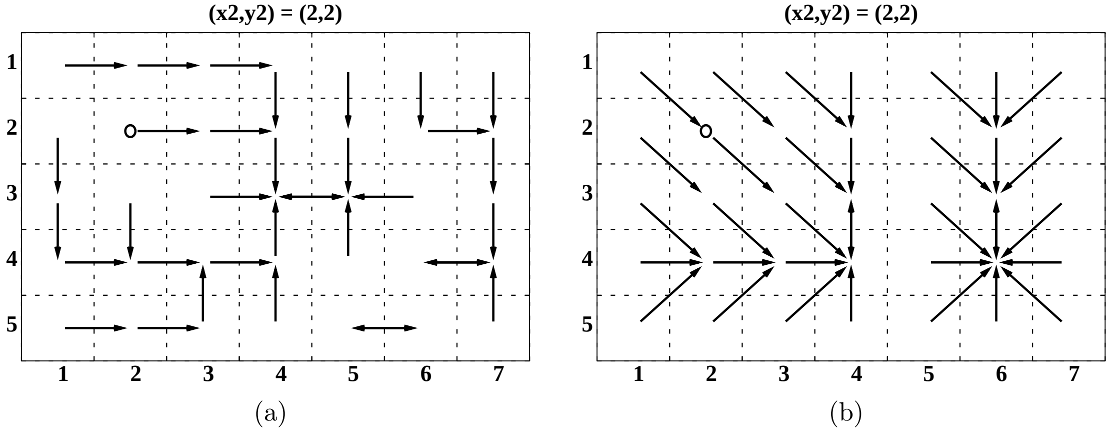


Figure 2: Visualization of two extracts of the computed policies: the uncontrolled player is marked by a circle and the arrows suggest the movement of the controlled player according to its position allowing four, respectively, eight moving directions.

movements – firstly, by applying no strategy to none of them, i.e., both walked randomly, and secondly, by equipping one player with the policy outlined above. The success of the strategy can be seen in the fallen ratio of collisions per moving steps from about 2.07% to mostly one or two accidents within the entire simulation. The few accidents that occurred are naturally explained by certain placements of the players on the square. In Figure 2(a), for instance, assume the controlled player sojourns in the upper left corner. As he/she has to move and all fields in reach are simultaneously reachable by the uncontrolled player as well, we cannot bypass a constellation that might lead into a collision. A closer inspection, however, reveals a correlation of the occurred collisions to the initial starting configuration of the players. As both of them have to move simultaneously and diagonal movings are prohibited, the *Manhattan distance*[1] between both players will either stay even or odd for the entire simulation. In the most crucial constellation of an odd Manhattan distance, i.e., if both are located face-to-face, the controlled player always has the possibility to move to at least one field that is out of reach of the uncontrolled player. Hence, and what our simulations have likewise shown, a collision can be ruled out if the Manhattan distance between both is odd. In case of a configuration with an even Manhattan distance, however, we have to face the aforementioned precarious situations in the corners of the square that might inescapably lead to a collision. A modification of our simulation solely ran with starting settings having an even Manhattan distance therefore yielded a collision ratio of 0.017%. One might argue, that, while neglecting initial settings, these situations could be avoided by an accurate policy. In any case, due to the undesirable influence of the initial configuration, we postpone this discussion after having adjusted the moving behavior of the actors that allows an alternating Manhattan distance characteristic.

[1] The Manhattan distance between two points A and B in the Euclidean space is computed over the sum of the absolute differences of their coordinates. In mathematical terms: Manhattan distance$(A, B) = |x_A - x_B| + |y_A - y_B|$.

Beside an extension of the action set $A' = A \cup \{$NORTH-EAST, SOUTH-EAST, SOUTH-WEST, NORTH-WEST$\}$ we also slightly adjusted the transition probability:

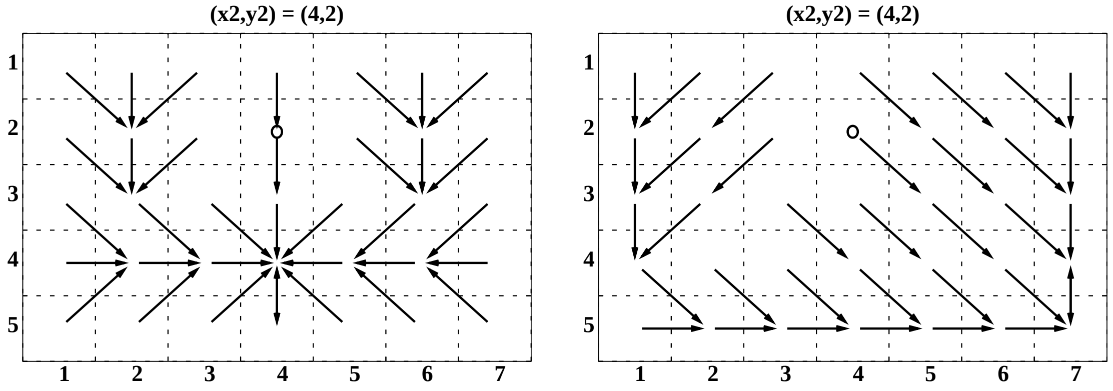
$$p'_{iaj} = \begin{cases} 0.33 & i \text{ comprises a corner field for the uncontrolled player} \\ 0.2 & i \text{ comprises an edge field for the uncontrolled player} \\ 0.125 & \text{otherwise} \end{cases}$$

Using these settings, we computed a new strategy (extract depicted in Figure 2(b)) that led to about 0.024% collisions in our simulations. Compared to the aforementioned scenario with an even Manhattan distance, the ratio slightly raised but simultaneously the number of precarious fields in the square increased from 4 to 20. Henceforth, not only the corner fields might lead into a collision but also constellations in which the controlled player resides on an edge field having the uncontrolled player right ahead. Consider, for instance, the following configuration in Figure 2(b): if the controlled player stays on the field $f_0 = (x, y) = (2, 1)$ all of the five possible succeeding fields are as well in reach by the uncontrolled player and, thus, might end up in a collision.

Again, the question raises whether a more accurate policy could prevent all or at least more of these precarious constellations. Intuitively, one might assume a smart strategy always tries to keep the largest possible distance to the uncontrolled player. For going into this matter, we adjusted the reward function in the following way: each or each possible collision is dealt in the same way as discussed before. Additionally, we weighted each system state with a *positive* reward derived from the Euclidean distance between both players.

$$r'_{ia} = r_{ia} + \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Therefore, the controlled player always prefers a field that enlarges or at least keeps the distance to the current position of the uncontrolled player. Exemplary, Figure 3 visualizes the policy without 3(a) and with 3(b) having modified the reward function for the same position of the uncontrolled player. Throughout all constellations we observed



(a)

(b)

Figure 3: Impact of the Euclidean distance to the reward function.

very similar suggestions of both strategies for crucial situations, i.e., if both players are very close to each other. In relaxed positionings, however, the policies significantly differ: by following its intention to enlarge the distance to the uncontrolled player, the policy considering the Euclidean distance tends to move the controlled player to the edges (cp. Figure 3(b)). But these fields, in fact, correspond to the dangerous situations that caused the observed accidents. Hence, our simulation resulted in an increased collision ratio from 0.023% to 0.09%.

2.3 Difficulties in VANETs

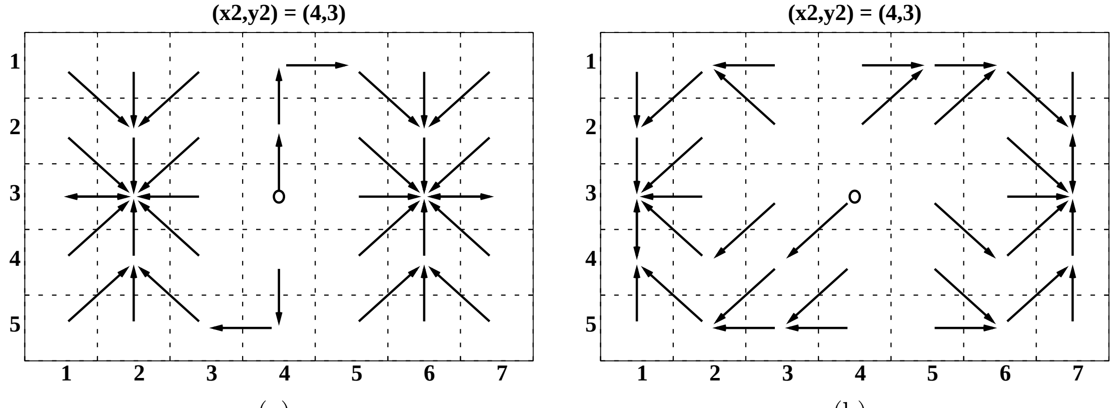
Up to now, the model presented so far does not show a huge affinity to a realistic vehicular scenario. Neither our players' moving characteristic matches a proper traffic representation nor does our model consider more than one uncontrolled players in the vicinity. In truth, a closer inspection of the particularities of VANETs brings up an additional problem that furthermore complicates the decision process. Due to adverse radio channel conditions our decision process needs to take a highly unreliable communication into account. Therefore, every protagonist does neither necessarily have a clear picture of his/her surrounding nor does he/she know with certainty whether the vicinity has received latest announcements. With reference to the above presented model, the latter problem could be handled by modifying the transition probability with which an actor goes from one state to another. In the following, let F denote the set of fields on the square and assume each player would only abruptly change his/her regular movement from one field to another – expressed by the function $\phi_{direction} : F \rightarrow F$ – on notification of a certain event by a not further specified instance. Further, let q_{trans} denote the probability for a successful transmission of a message. Hence, from the point of view of player 1 that has just received a relevant message and chose $a \in A$ as reaction on that incident, the system will segue from the current state $i = (i_1, i_2) \in S$ to the future state $j = (j_1, j_2) \in S$ with the following probability ($i_1, i_2, j_1, j_2 \in F$ denote occupied fields by player 1 or 2, respectively):

$$p'_{iaj} = \begin{cases} q_{trans} \cdot p_{iaj} & \text{if } \phi(i_2) \neq j_2 \\ (1 - q_{trans}) \cdot p_{iaj} & \text{if } \phi(i_2) = j_2 \end{cases}$$

In a random walk scenario as presented above we will hardly find a useful regular movement function ϕ as each direction is chosen with the same probability. After having exchanged the moving behavior with an existing traffic model, however, we expect to find regularities in the movement of vehicular that allows an appropriate supposition on the movement. In the following we address the first mentioned problem of a possible imperfect view on the vicinity.

In VANETs periodic messages are often conceived to keep an up-to-date view on the traffic situation in the close surrounding. By means of this knowledge a Driving Assistant System could warn a driver of taking precarious actions or consult him/her in order to manage critical situations. An unreliable communication between cars, however, could lead to a loss of accuracy concerning the assumption of the location of other vehicles. Consequently, a Driving Assistant System can only estimate the current system state and needs to find the best possible action on basis of this guess.

We refined the model presented in Section 2.2 and blurred the actual position of the uncontrolled player by means of a cloud that additionally covers each adjacent field (i.e. maximal 9). Due to simplicity, we assumed the uncontrolled player to be uniformly distributed within the cloud. The adjusted simulation yielded a collision probability of 0.045%. Following the same argumentation that explained the disadvantageous influence of the Euclidean distance to the reward function, naturally, the increased ratio of collision is caused by the tendency to move the controlled player to the edges. As the evolution from Figure 4(a) to Figure 4(b) shows, the number of eluding fields shrinks and these fields themselves get edged out to the borders of the square. Naturally, this evokes the aforementioned improper settings that might lead to a collision.



(a)

(b)

Figure 4: Visualization of strategies assuming exact and blurred location of counterpart.

The uncertainty due to the unreliable communication modeled above, covers 9 out of 35 fields, i.e. more than a fourth. Therefore, this might be a representation for scenarios in which a highly unreliable transmission is expected. A variation of the size of the cloud easily adjusts the model to the environment. Figure 5 compares the degree of uncertainty, i.e. the size of the cloud, to the expected ratio of collisions.

2.4 Remarks

Before we face the problem from the perspective of the communication system we list some notes concerning difficulties with the toy example and concerning the estimation of our results.

The computation of an optimal policy used in the toy example above requires a huge effort that does not meet restrictions of real-time applications. In (Dolgov and Laberteaux, 2005) the authors addressed the problem and proposed linear approximations to the Linear Program in order to reduce the mathematical complexity in terms of magnitudes. The trouble with this approach, however, stems from the difficulty of finding appropriate linear approximations. In other words, it is hard to choose a

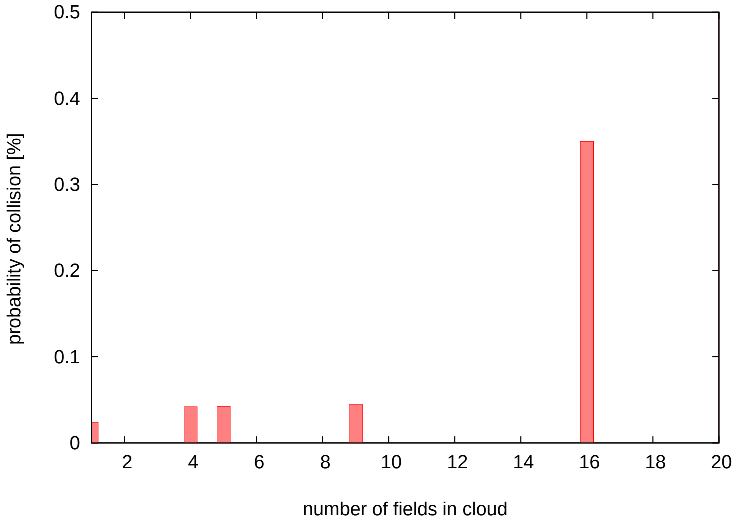


Figure 5: Effect of uncertainty, expressed by the size of the cloud, to the probability of transmission.

relaxation that generates a policy with a comparable success as if the entire system had been taken into consideration.

Although, our simulations have shown the impact of uncertainty to the accuracy of a Driving Assistant System, the reader should be aware that this effect gets even more stressed when dealing with more than two players in the scenario. Assume the aforementioned nine-cloud scenario but more than one uncontrolled player and the worst case of disjunct uncertainty-clouds. The probability of guessing the correct current system state, $P(\text{current state} = i) = \left(\frac{1}{9}\right)^n$, falls for $n = 1$ from approximately 11% to about 0.14% for $n = 3$. As a trailing effect, we expect the ratio of collisions to increase likewise exponentially which easily exceeds the shown effect in Figure 5.

3 Optimizing the Probability of Reception

In Section 2.3 we brought up the problem of unreliable communications in VANETs that, especially in situations of acute danger, might diminish the usefulness of any Driver Assistant System. The belief in a life saving, or at least accident preventing, system could hardly be preserved if it only works in some out of all critical cases. Therefore, we must ensure the communication system to provide a certain reliability. Several rebroadcasts could remedy as they raise the transmission probability but likewise increase the load on the communication channel what might, in turn, impede the dissemination of any other urgent message. Again, we run into an optimization problem that needs to ponder the load on the communication channel on the one hand and an advisable transmission probability on the other hand. The contribution of this section is the formalization of the according optimization problem.

From an abstract point of view, we deal with an originator, a node that initiated the data dissemination, and several receiving nodes within a given area. Due to the vehicular application, we can restrict the area of interest to street passages in the surrounding of the originator. Depending on their position to the originator, we assume the receiving nodes to have a varying need for a reliable reception of the message: exemplarily, nodes closer to the originator might be more reliant on the information than nodes at a farther distance. As the initial transmission by the originator might not reach every node in the area of interest or a successful reception is only expected with a low probability, the likelihood of reception could be locally increased by letting receiving nodes rebroadcast the message. Naturally, the effect of a retransmission depends *i)* on the probability with which the resending node has received the message before, *ii)* on the number of nodes that would benefit from this retransmission and *iii)* on interferences with other, simultaneous retransmissions impeding a successful transmission. We call a coordination of rebroadcasts, a convention which subset of nodes retransmits the message at which point in time, a *broadcast sequence.* The success of a broadcast sequence can exemplarily be measured by the time or the number of rebroadcasts that are required to achieve a given probability of reception.

What we intend to find is a broadcast sequence that achieves given geographic-depending probabilities of reception with a minimum amount of retransmissions. In order to solve this problem, again, by means of a Markov Decision Process we consider a part of a street that we divide into stripes numbered from 1 through n (cp. Figure 6). Each receiving vehicle determines its corresponding stripe by means of GPS-data in the alert message. Let $q_i \in Q$ denote the current reception probability of vehicles in stripe i with Q being, due to simplicity, a discrete set of probabilities from 0% through 100%. The temporary probability of reception q_i might suffice the initial given demand, R_i, ($q_i \geq R_i$) or is reliant on rebroadcasts of vehicles in other stripes in order to do so ($q_i < R_i$). As each of the stripes could have received the message with $|Q|$ possibilities, the entire regarded system comprises $|S| = |Q|^n$ states. The complexity of the mathematical problem becomes furthermore stressed as in each of the states, every subset of the stripes might rebroadcast the alert simultaneously, i.e., the amount of actions accounts $|A| = 2^n$. The transition probability function involves another problem: investigations have shown that in absence of interferences the probability of reception decreases with increasing distance to the sender according to the *Nakagami Radio Propagation Model* (Torrent-Moreno, Killat, and Hartenstein, 2005). Hence, a

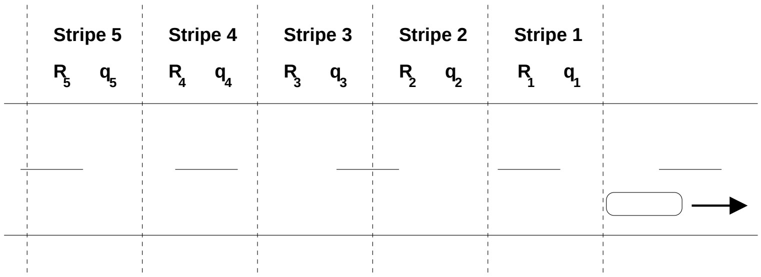


Figure 6: Dividing the street into stripes in order to find an appropriate broadcast sequence.

successful reception within a stripe depends, on the one hand, on the distance to the sender and, on the other hand, on concurrent other senders whose transmissions might interfere and destroy the packets. Simultaneous rebroadcasts of neighboring stripes, for instance, solely yield little benefit as most of both transmission signals are destroyed anyway. Our goal to achieve the initial postulated stripe-individual probability of reception, R_i, with a minimum amount of rebroadcasts is taken into account by the reward function that simply punishes each rebroadcast with a negative value.

Clearly, the problem asks even more for a drastic mathematical relaxation as the exponential complexity cannot only be ascribed to the number of system states but also to the size of possible actions. Moreover, note that the above sketched simplification assumed a vehicle to sojourn in each of the stripes for what a once computed broadcast sequence would suffice. In reality, however, we must deviate from such an assumption and compute an accurate flooding sequence depending on the density of traffic on the street.

4 Conclusion

In this paper we outlined the problem of keeping a strict separation of communication and application layer by means of a collision avoidance application in Vehicular Ad-hoc Networks. Firstly, we have shown on basis of a drastic simplification, how unreliable communication impairs a proper decision process and results in an increased number of accidents. Secondly, our attempt to counter the effect by improving the reliability of communication got impeded by real-time application constraints. By taking a 'super-position' that simultaneously considers the requirements of the application and the constraints of the communication system (as indicated in Section 2.3), however, we are convinced to increase our global objective 'safety of driving'. A quite comparable situation face agents in Market Engineering as they optimize the global objective by adhering to the market underlying rules.

Acknowledgment

We would like to thank Dmitri Dolgov and Ken Laberteaux from Toyota Technical Center, Ann Arbor, USA, for their work on the formulation of the MDP problem and for their comments on an earlier version of this paper.

References

Dolgov, D. A. and K. Laberteaux (2005): "Efficient Linear Approximations to Stochastic Vehicular Collision-Avoidance Problems," in: *Proceedings of the Second International Conference on Informatics in Control, Automation, and Robotics (ICINCO-05).*

Puterman, M. L. (2005): *Markov Decision Processes*, Wiley Interscience.

Torrent-Moreno, M., M. Killat, and H. Hartenstein (2005): "The Challenges of Robust Inter-Vehicle Communications," in: *Proceedings of the 62nd IEEE Semiannual Vehicular Technology Conference (VTC-Fall).*

A Markov Decision Process

A Markov Decision Process (MDP) as we use it in our context is defined as a 5-tuple $< S, A, p, r, \gamma >$ with the following interpretation (cp. (Puterman, 2005) and (Dolgov and Laberteaux, 2005)):

S: A finite set of states describing the entire system.

A: A finite set of actions that an actor can execute.

p: $S \times A \times S \to [0,1] : p_{iaj}$
A function determining the transition probability of going to system state $j \in S$ when taking action $a \in A$ in system state $i \in S$. Furthermore, we assume p to be stochastic, i.e., $\sum_{j \in S} p_{iaj} = 1$.

r: $S \times A \to \mathbb{R}$: r_{ia}
A function determining the one-stage reward that an actor obtains for taking action $a \in A$ in state $i \in S$.

$\gamma \in (0,1)$: Discount factor

As we conceive the MDP to be discrete we are dealing with a clear separation of time periods in which an agents observes the current state of the system and carries out an action. Depending on the chosen action and on the current state, the agent obtains a reward that is expressed by the function r. Likewise, the succeeding system state is determined by the transition law p.

In addition to stationary policies $\pi = (f_0, f_1, \ldots)$ or simply f (cp. section 2.1) we may also apply policies π depending on the history of the process and/or policies π with actions randomly chosen.

It is well known, however, that for each such policy π there always exists some stationary policy f with an improved expected total discounted reward, i.e., $v_f(i) \geq v_\pi(i)$ for all $i \in S$.

Competing on Many Fronts: Entry Networks in an Economy of Multi-Product Firms

Jan Krämer[1], Siegfried Berninghaus[1], and Christof Weinhardt[2]

[1] Institute for Economic Theory and Operations Research,
Universität Karlsruhe (TH)
{kraemer,berninghaus}@wiwi.uni-karlsruhe.de

[2] Institute of Information Systems and Management,
Universität Karlsruhe (TH)
weinhardt@iw.uni-karlsruhe.de

Summary. We consider an economy with n distinct firms, each of which has a unique set of core competencies and thus initially serves its home product market as a monopolist. We then allow for informational spillovers, enabling firms to acquire foreign competencies in order to enter other product markets. Upon entry a firm has to invest sunk costs and is, of course, only willing to do so if it believes that post-entry profits will exceed the entry costs. Conversely, an entrant's home market may also be threatened by entry. This assumption is new to the industrial organization literature and gives - among other effects - rise to a prisoners' dilemma. As a consequence, we can formally establish the result that firms tend to overcompete when operating on many markets. Thus, one implication of the model is that multi-market competition has positive effects on consumers' surplus.

1 Introduction

Traditionally, game-theoretical works on market entry have merely considered settings in which a single incumbent is faced by a single entrant on exactly one market. Although this basic setting is well understood today and has been looked at from many different angles[1], surprisingly, with some notable exceptions, game theorists have thus far barely strayed from the very limited scope of this model. More specifically, present models largely neglect settings with multiple markets or ignore the fact that entrants themselves might indeed be an incumbent elsewhere. In this paper we make a first effort in addressing this lacuna in the literature by proposing a market entry model with multi-product firms, each of which has a home market.

To this end, each firm in our model is considered to have a unique set of *core competencies*, which are initially utilized for the production of a single good, called the *core product*. We use the term of core competencies very broadly, comprising e.g. knowledge

[1] In particular for varying informational properties. See e.g. Neven (1989) and Wilson (1992) for excellent surveys.

of production processes, raw materials, supply and distribution channels, results from R&D, such as patents and process innovation, information about customers and their behavior on the home market, as well as informal and formal social ties facilitating production and distribution. Endowed with such a unique set of special skills, each company produces and supplies a single distinct good as a monopolist on her *home market*.[2] Thus, in the absence of competency spillover, we envision a set of unrelated (home) product markets, each of which is supplied by exactly one firm in the economy. In reality, however, spillover effects (whose various sources include through publicly observable information such as patents or market prices, tacit knowledge of employees, learning effects or even espionage) cannot be neglected. Thus, in the long run, firms are not limited to the production of their core product and may chose to imitate other firms' products. In our model, this assumption bears two main implications:

Firstly, in order to produce a foreign product, firms have to acquire at least some of the respective competitor's competencies in order to be able to produce the same commodity.[3] It is reasonable to assume that the informational spillover between firms is not perfect; therefore the production of a foreign good is c.p. more demanding than production of the core product for any given firm. For example, a company manufacturing ships might, in principle, also be able to build cars, but lack of experience in this area certainly makes it much harder to match a competitor who has been making cars from the start. In order to model this formally, we need to quantify exactly how much harder it is for a firm to imitate a competitor's product. We therefore introduce a *technological distance*, which measures the relatedness between core products, or - more generally - between the set of core competencies belonging to the respective firms. At this juncture, it is not necessary to elaborate on the precise nature of this measure because it is not crucial for understanding the model. For the time being, technological distance can be thought of as a scalar measure.

Secondly, firms must enter foreign markets in order to sell their new product. Here we follow traditional concepts and assume that whenever a firm decides to enter a distant market, it has to bear sunk costs of market entry. In this context, entry costs may consist of advertisement costs, changeover-times or even procurement of new machinery so specific to the cause of the new market that it cannot be resold. In particular, the acquisition of core competencies with respect to the foreign good or market are also considered sunk costs, whose magnitude in turn depends on the technological distance. However, we would again like to point out that spillovers are imperfect. The incumbent firm will thus always be able to maintain her informational head start and produce her core product more cost-effectively.[4] Of course, a firm is only willing to sacrifice sunk costs in a new product market if it deems the potential post-entry profits sufficient to

[2] In a single market context we use the convention of addressing an incumbent as female and an entrant as male.

[3] By foreign or distant goods (markets) of a firm, we denote all goods (markets) except for the core product (home market) of that firm.

[4] An alternative argumentation would be that the incumbent firms will always produce a good of higher quality. Since we assume homogeneous product markets, however, this interpretation is not directly applicable here.

recoup these expenditures. In order to mirror the informational head start by the incumbent, we chose to model post-entry competition by an asymmetric Cournot game, where the home firm has a comparative cost advantage over her competitors.

This framework allows for a multitude of investigations, and we will present some extensions to the basic model later. However, the quest for stable and efficient constellations for the multi-product firm economy constitutes the focus of this paper. In other words, for a given economy, we will show which firms produce which goods in equilibrium and whether this equilibrium maximizes total welfare. In addition, we will also explore the causes of any existing inefficiencies. To this end, we first present a formal description of the model in Section 2. Section 3 subsequently reveals insights on the precise nature of post-entry competition. Section 4 introduces the concept of entry networks and develops a complete characterization of stable and efficient entry networks. Section 5 elaborates on the basic tension between stable and efficient outcomes; the paper concludes with Section 6.

2 The Model

Let $N = \{1, \ldots, i, \ldots, n\}$ denote a set of firms. Each firm has a home market for which it produces a homogeneous good. Thus, in total, n distinct markets exist in the economy.[5] These markets are assumed to be arranged spatially according to some technological distance, d, defined on the n products.[6] More specifically, by d_i^j we denote the distance between product i and j, where d_i^i is normalized to $d_i^i = 0$. In principle, each firm i can produce any of the n goods at costs of $c_i^j(q_i, d_i^j)$, where q_i denotes i's output quantity. However, it is implicitly assumed that it is more costly to produce distant goods, i.e.

$$c_i^r(q_i, d_i^r) \geq c_i^s(q_i, d_i^s) \quad \forall i, q_i, \text{ if } d_i^r > d_i^s. \tag{1}$$

If firm i seeks to enter market $j \neq i$, it has to bear entry costs of f_j.[7] Upon entry in market j, firm i competes against $n_j - 1$ other firms in that market. Hence, depending on i's cost structure, its strategy, and the demand, D_j, it receives net payoffs of

$$\Pi_{i,j} - f_j \tag{2}$$

for each market j in which it competes. Furthermore, it is naturally assumed that $\Pi_{i,j}$ is decreasing in the number of competitors in market j, i.e.

$$\frac{\partial \Pi_{i,j}}{\partial n_j} \leq 0. \tag{3}$$

[5] Since there is a one-to-one mapping between markets and their respective home incumbents, we use the same indices to label markets and firms. However, generally, we will denote a representative firm by i and a representative market by j.

[6] Although distance may be vectorial, we will from now on consider a scalar measure. This does not limit generality, however, since a scalar can always be deduced from a vector by means of a metric, such as the euclidean distance.

[7] Note that f itself may depend on a variety of variables, such as the distance d, or the current competition in the market. Moreover, for the home market, i, it is assumed that $f_i = 0$ holds.

Of course, firm i will only wish to enter market j, if (2) is strictly positive. Thus, in order to make the setting tractable, one must propose a specific cost- and demand-function along with the precise nature of post-entry competition. Here we assume that in each market firms set quantities in a Cournot game. Furthermore, all markets are identical with respect to an inverse demand of

$$D_j^{-1} = p(Q_j) = a - Q_j, \tag{4}$$

where Q_j denotes the aggregate output produced by the n_j firms present in market j. In addition, firms are also symmetric, each having a cost function of $c_i(q_i, d_i^j) = (c + d_i^j)q_i$ with $c \geq 0$. This function reflects the imperfectness of informational spillover and grants comparative cost advantages to the incumbent firm, since her marginal costs reduce to c.[8] Hence, in her home market, the incumbent firm's profit is always higher than that of her competitors and ensures that she will never exit it. Of course, by this we implicitly assume that every market in the economy is profitable to a monopolist. Moreover, for simplicity, in the following we will not differentiate between distant competitors, thus setting $d_i^j = d$, for $i \neq j$. Finally, we assert that the market entry cost, f, is exogenously given and uniform across markets and firms. Certainly, such concrete assumptions will always fall prey to criticism; however, we believe to have sufficiently motivated our choices and are confident that the subsequent results generalize to more complex settings.

3 Post-Entry Competition

From (2) it is obvious that a firm's decision whether or not to enter a foreign market depends solely on two factors: the payoffs from post-entry competition and the amount of sunk costs it has to bear. In this section we will have a closer look at the per market payoffs of the home incumbent and the foreign competitors. More precisely, in market j there are a total of n_j competitors, one home incumbent and $n_j - 1$ competitors from distant markets. In equilibrium, each firm chooses its output q_i^* so that

$$\max_{q_i} \ \mathit{\Pi}_i(q_i, q_{-i}^*) = \left[a - q_i - \left(\sum_{j \neq i} q_j^* \right) \right] q_i - c_i q_i \tag{5}$$

which, assuming $q_i^* \geq 0$ for all i, yields n_j different first-order conditions

$$a - 2q_i^* - \left(\sum_{j \neq i} q_j^* \right) = c_i, \quad i = 1, \ldots, n_j. \tag{6}$$

Instead of solving n_j equations for n_j output levels, we solve for the aggregate production level by rewriting (6) to

[8] The reader can easily verify that the properties (1) and (3) are met.

$$a - q_i^* - Q^* = c_i, \quad i = 1, \dots, n_j. \tag{7}$$

Summing over all $q_i, i = 1, \dots, n_j$ yields

$$n_j a - Q^* - n_j Q^* = \sum_{i=1}^{n_j} c_i \tag{8}$$

Hence, the Cournot equilibrium aggregate industry output and market price are given by

$$Q^* = \frac{n_j a - \sum_{i=1}^{n_j} c_i}{(n_j + 1)} \tag{9}$$

$$p^* = \frac{a + \sum_{i=1}^{n_j} c_i}{n_j + 1}. \tag{10}$$

Notice that under constant unit costs, industry output, price, and therefore total welfare can be calculated using the sum of the firms' unit costs without investigating the precise cost distribution among firms. Moreover, this result does not rely on linear demand and therefore also applies to nonlinear demand functions. However, in our setting, there are only two distinct types of firms. Namely, the home incumbent having constant unit costs of c and the foreign entrants having constant unit costs of $c+d$. Thus (9) reduces to

$$Q^* = \frac{n_j(a - c) + (n_j - 1)d}{n_j + 1} \tag{11}$$

$$p^* = \frac{a + n_j \; c + (n_j - 1)d}{n_j + 1} \tag{12}$$

and from (7) we can calculate q_i^* and finally the profit functions $\Pi_i^*, i \in \{H, F\}$ for the home incumbent (H) and the foreign entrant (F), respectively:[9]

$$\Pi_H^* = \left(\frac{a - c - (n_j - 1)d}{n_j + 1}\right)^2 \tag{13}$$

$$\Pi_F^* = \left(\frac{a - c - 2d}{n_j + 1}\right)^2 \tag{14}$$

We set $\gamma := a - c$, such that (14) can be rewritten as $\Pi_F^* = \left(\frac{\gamma - 2d}{n_j + 1}\right)^2$. Moreover, since $q_i > 0$ must hold in equilibrium, it follows that $\gamma > 2d$. Furthermore, for a given γ and d we adopt the shorthand notation $\Pi_i^{n_j}$ to denote the equilibrium payoff of firm $i \in \{H, F\}$ in a market j where n_j competitors are present.

[9] Note that $\Pi_i^* = (q_i^*)^2$

4 Entry Networks

In this section we establish a graphical representation of the entry decisions of the n firms that greatly helps to facilitate the future analysis and will, at a later time, visualize the interdependencies between markets and firms. To this end, we represent each market (and thus each home incumbent) by a graph vertex and label these accordingly by $N = \{1, \ldots, i, \ldots, n\}$.[10] We then draw a directed edge from vertex i to vertex j iff firm i has entered market j. The resulting equilibrium network is called the *competitive entry network* of the economy. Notice that the competitive entry network represents not only the entry decisions of each firm (and thus its product portfolio), but also the fierceness of competition on each market.[11] Likewise, the *efficient entry network* represents a constellation of the economy which maximizes total welfare. However, neither the efficient nor the competitive entry network has to be unique.[12] The representation of market entry decisions via a network is new to the literature and particularly well suited to study cross market (or cross firm) effects. In this vein, the present model provides a framework that bridges traditional models of industrial organization with recent works on strategic network formation (see e.g. Bala and Goyal, 2000). At present, only very few models share this feature (e.g. Goyal and Moraga-Gonzalez, 2001; Goyal and Joshi, 2003; Billand and Bravard, 2004), but none of them deals with market entry. Furthermore, in contrast to classic models of market entry, here entrants are deanonymized and have a "life outside of the focal market". We will briefly return to this point later, but unfortunately, due to restrictions in space, we cannot elaborate on this important and interesting facet of the model. In what follows we will therefore restrict ourselves to a complete characterization of the efficient and competitive entry networks.

4.1 Competitive Entry Networks

We consider myopic firms, which is a common assumption in network formation and reflects the fact that firms cannot boundlessly foresee the consequences of their actions due to the complexity of the game. Thus, firm i will seek to enter distant market j, iff

$$\Pi_F^{n_j} = \left(\frac{\gamma - 2d}{n_j + 1}\right)^2 > f \tag{15}$$

Since we have assumed symmetric firms and markets, a simple argument shows that for each competitive entry network, each market must indeed possess the same number of competitors. It follows from (15) that the number of competitors in equilibrium, denoted by $\eta = n_j$, $\forall j = 1 \ldots n$ solely depends on the market entry costs, f. However, η does not relate to the particular shape of the competitive entry network and is therefore merely a measure of the competitiveness of the individual markets, without

[10] Moreover, the vertices may be arranged in such a way that the distance between them reflects the technological distance between the respective core products of the markets.

[11] This follows directly from (3).

[12] We will show the latter by example in Section 5.

revealing precise information about a firm's individual payoffs. To see why η must be constant across markets, consider a particular market, say j, with only $n_j < \eta$ competitors in equilibrium, while on all other markets there are η competitors. Then, of course, it would be profitable for one foreign competitor $i \neq j$ to enter market j, since by assumption f, $D^{-1}(p)$ and the post-entry competition payoffs are identical for all markets. Likewise, if more than η firms compete in a specific market, at least one firm must receive negative overall payoffs, i.e. $\Pi_F^{\eta} - f < 0$, and would thus be better off by not entering that market. In summary, we obtain that in equilibrium

$$\eta = \max_{k \geq 1} \{k \leq n | \Pi_F^k \geq f\} \tag{16}$$

Consequently, since in equilibrium there are $\eta - 1$ foreign entrants in every market, the competitive entry networks must satisfy

$$\delta_i^{in} = \eta - 1, \qquad \forall i = 1 \ldots n, \tag{17}$$

where δ_i^{in} is the in-degree of vertex i in an entry network. From now on we will say that an entry network which satisfies (17) is an η-network. Furthermore, from (16) we can directly follow that the economy switches from an η-network to an $(\eta - 1)$-network as soon as f rises just above Π_F^{η}. We denote this by

$$\eta \rightarrow \eta - 1 \qquad \text{at} \qquad f = \Pi_F^{\eta}. \tag{18}$$

Notice that (17) does not determine a unique network, and among the variety of equilibria, there also exist some which are asymmetric. More explicitly, this means that in spite of ex-ante symmetric firms and markets, there exist scenarios in which some firms are able to extract greater rents from the economy at the cost of others.

4.2 Efficient Entry Networks

Having established the necessary and sufficient condition for competitive entry networks, we would now like to explore whether these networks are efficient, and, if not, what the properties of efficient networks are. Given the above model of Cournot competition, consumers' surplus per market is given by:

$$CS_{n_j} = \frac{(a - p_i)^2}{2} = \frac{(\gamma - d)(n_j - 1) + \gamma}{2(n_j + 1)^2}. \tag{19}$$

Obviously, $\frac{\partial CS_{n_j}}{\partial n_j} > 0$, i.e. consumers' surplus increases with competition. Likewise, consumers' welfare decreases with the distance between markets, since $\frac{\partial CS_{n_j}}{\partial d} < 0$. Furthermore, producers' surplus per market amounts to

$$PS_{n_j} = \Pi_H^{n_j} + (n_j - 1)\ \Pi_F^{n_j} \tag{20}$$

and one can easily verify that $\frac{\partial PS_{n_j}}{\partial n_j} < 0$. Thus, total welfare per market, given by $W_{n_j} = CS_{n_j} + PS_{n_j}$, can be written as follows:

$$W_{n_j} = \frac{\gamma^2 n_j(n_j+2) - 2d\gamma(n_j^2+n_j-2) + d^2(3n_j^2+2n_j-5)}{2(n_j+1)^2} \tag{21}$$

and increases with competition, i.e. $\frac{\partial W_{n_j}}{\partial n_j} > 0$. We can once more exploit the symmetry assumption to derive that $\eta = n_j$, $\forall j = 1 \dots n$ must hold once again.[13] It is then obvious that the overall welfare of the economy depends only on η and the entry costs, f:

$$W^{\eta} = n * W_{\eta} - (\eta - 1)nf \tag{22}$$

Formula (22) reveals the two opposing forces of welfare maximization. The first term, relating to per market welfare, has been shown to increase in η. On the contrary, the second term, representing the total amount of sunk costs invested by all entrants, decreases in η. Hence, determination of the efficient entry networks is non-trivial. However, since n is a constant factor in both terms, it is not relevant for welfare maximization, and we obtain:

$$W^{\eta} \rightarrow W^{\eta-1} \qquad \text{at} \qquad f = W_{\eta} - W_{\eta-1}, \tag{23}$$

where $W_{\eta} - W_{\eta-1} = \frac{\gamma^2(1+2\eta) - 2d\gamma(\eta^2+5\eta+2) + 4d^2(\eta^2+3\eta+1)}{2\eta^2(\eta+1)^2}$.
Having established both the efficient and the competitive transition point of the economy, we now turn to a comparison and interpretation of the two in the next section.

5 Stability vs. Efficiency

From formula (23) we know at what point it is efficient for the economy to switch from an η- to an $(\eta-1)$-network. Likewise, formula (18) specifies where this switch will occur under competition. Thus, by subtracting the efficient transition point given by (23) from the competitive switching point in (18) we obtain a simple measure of the inefficiency in the economy, denoted by $\varDelta W^{\eta}$:

$$\varDelta W^{\eta} = \frac{\gamma^2(2\eta^2 - 2\eta - 1) - 2d\gamma(3\eta^2 - 5\eta - 2) + 4d^2(\eta^2 - 3\eta - 1)}{2\eta^2(\eta+1)^2} \tag{24}$$

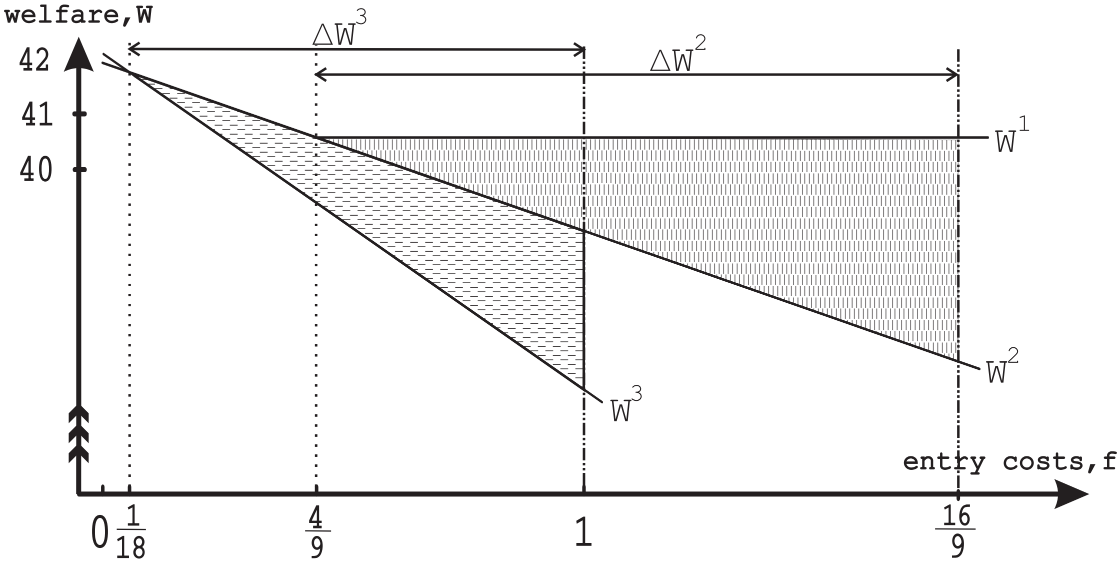
In the relevant parameter range $\varDelta W^{\eta}$ is strictly positive,[14] which means that firms switch too late from an η to an $(\eta-1)$ network. Thus, under the assumptions of our model, multi-market entry results in excessive competition and therefore increases consumers' welfare. Moreover, since $\frac{\partial \varDelta W^{\eta}}{\partial d} < 0$, we obtain the interesting result that the

[13] With a slight abuse of notation, we denote by η the number of competitors per market in both the efficient and competitive entry networks.

[14] Recall that $\gamma > 2d$ and, obviously, $\eta - 1 \geq 1$. All (following) results are conditional to these properties.

inefficiency is mitigated by an increase in the imperfection of informational spillovers.[15] Moreover, for completeness, we state that the welfare losses decrease as the competition, i.e. η, increases on each market.

Now, for expositional clarity, consider an economy consisting of $n = 3$ firms and let the imperfectness of informational spillover be $d = 1$. Furthermore, we set $\gamma = 6$, such that Π_F^3 is normalized to one. Figure 1 graphically depicts the efficient and competitive transitions in this economy, as derived by the formulas (18) and (23). The f-values at which the efficient transitions occur are indicated by the two dotted vertical lines, where the first line at $\frac{1}{18}f$ represents the switch from the 3-network to a 2-network, and the second line at $\frac{4}{9}f$ denotes the switch from a 2-network to the 1-network. Likewise, the corresponding competitive transitions are given by the two vertical dashed and dotted lines at f and $\frac{16}{9}f$, respectively. In addition, this is also expressed in the table below the graph, where the competitive and efficient network-structures are given in terms of η. It is easy to see that the efficient transitions indeed occur much later, as expressed by the positive values for both $\varDelta W^3$ and $\varDelta W^2$. In fact, with the visual aid of Figure 1, it is obvious that the total inefficiency in the economy is precisely given by the shaded area between the economy's efficient fringe and the corresponding welfare functions W^η. Thus, formally, total inefficiency can be calculated by



η (efficient entry network)	3	2	1	1	1
η (competitive entry network)	3	3	3	2	1

Figure 1: Efficient and Competitive Transitions in the Example Economy

[15] An interpretation of this result is postponed until the end of this section.

$$\Delta W = \sum_{\eta=2}^{\eta=n} \left| \int_{\max\{0, W_\eta - W_{\eta-1}\}}^{\Pi_F^\eta} \left(W^\eta - W^{\eta-1} \right) df \right| . \tag{25}$$

In particular, notice that the difference between the upper and lower limit of the integral is exactly ΔW^η.[16] In Figure 1, the area represented by the first summand of (25) has been shaded by '|'s, whereas the area corresponding to the second summand has been shaded by '−'s. Total inefficiency amounts to $\frac{8}{3} + \frac{289}{216} \approx 4$. To see that the welfare loss is indeed reduced if the informational spillover decreases, consider the same economy, but with $d = 2$. The reader may verify that the corresponding welfare functions are $W^1 = \frac{81}{2}$, $W^2 = \frac{118}{3} - 3f$, and $W^3 = \frac{309}{8} - 6f$ and that the competitive transitions occur at $\Pi_F^3 = \frac{1}{4}$ and $\Pi_F^2 = \frac{4}{9}$. Thus, equation (25) yields an overall welfare loss of $\frac{22}{27} + \frac{13}{48} \approx 1.1$, which is significantly smaller than in the economy where $d = 1$.[17]

Next, we investigate the competitive entry networks for the example economy. From equation (17) we know that at the extremes, i.e. for very small ($f \leq \Pi_L^3$) and for very large ($f > \Pi_L^2$) values of f, the competitive entry networks are uniquely determined to be the complete and the empty network, respectively. For intermediate values of f, however, the equilibrium network is not unique. In particular, in this example both network architectures depicted in Figure 2 satisfy condition (17) for $\eta = 2$.[18]

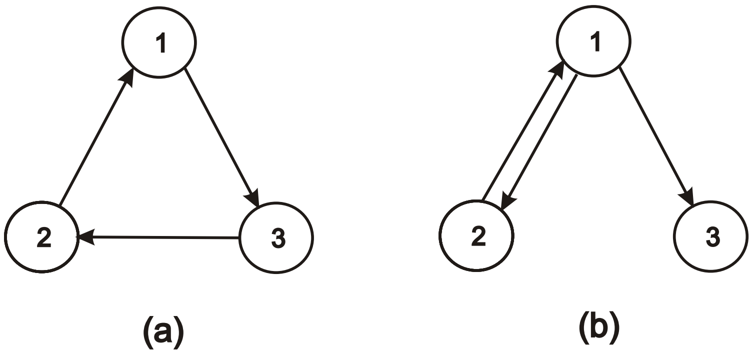


Figure 2: Some Competitive Entry Networks in the Example Economy

The graph from Figure 2(a) is perfectly symmetric (circular). Each firm invades exactly one distinct foreign market and its home market is in turn invaded by exactly one other foreign competitor. Thus, firms' payoffs are also symmetric, amounting to $\Pi_H^2 + \Pi_F^2 - f$ for each firm $i \in \{1, 2, 3\}$. In general, the symmetric payoff in an η-network ($\eta \geq 2$) is

$$\Pi_H^\eta + (\eta - 1)(\Pi_F^\eta - f). \tag{26}$$

[16] Under the assumption that $W_\eta - W_{\eta-1} > 0$, which is generally true for $d < \frac{\gamma(1+2\eta)}{2(\eta^2+3\eta+1)}$.

[17] Also notice that in this economy the complete 3-network is the unique efficient network for all feasible values of f.

[18] A network architecture is a set of graphs that differ only with respect to the labeling of their nodes.

Notice that interestingly, the total payoff to firm i is well below the payoff it would receive if each firm were to operate in her respective home market only. To see this, recall that Cournot competition is not Pareto optimal and hence $\Pi_i^1 > \eta \Pi_i^\eta$, $i \in \{H, F\}$ holds. Trivially, since $\Pi_F^\eta < \Pi_H^\eta$ the result obtains, even without considering entry costs. The firms thus find themselves in a prisoners' dilemma, in which it is collectively rational (and efficient!) to stay in isolated monopoly markets, but where the nature of the entry-game forces the firms into excessive competition.

Figure 2(b) shows the second equilibrium network of the example economy, which yields asymmetric payoffs to the ex-ante symmetric firms. Here firm 1 invades both, market 2 and market 3, leaving only her own home market profitable enough for entry by firm 2. Firm 1 certainly earns the highest profits in the economy, amounting to $\Pi_H^2 + 2\Pi_F^2 - 2f$. More generally, in any η-network ($\eta \geq 2$), the highest profits an individual firm, the *predating firm*, can gain are

$$\Pi_H^\eta + (n-1)(\Pi_F^\eta - f). \tag{27}$$

In this scenario we cannot generally observe a pure prisoners' dilemma. Whether the predating firm's profits are above or below the monopoly level depends on two factors: the size of the economy, n, and the profitability of foreign entry. The latter cannot exceed $\Pi_F^\eta - \Pi_F^{\eta+1}$ in the relevant parameter range for f, since otherwise the economy would switch from an η-network to an $(\eta+1)$-network, contradicting stability. Furthermore, from (26) we know that $n \gg \eta$ must hold if the predating firm is supposed to gain more profits than in the monopoly 1-network. In principle, the model certainly allows for the possibility that an individual firm extracts more profits from entering foreign markets than by remaining in isolation and maintaining a monopoly market. However, this possibility is rather theoretical in nature, since it requires either that the predating firm is capable of invading a large number of markets at once or that her competitors stay passive while being predated. Although the model presumes an ad-hoc equilibrium at this point and does not put any restrictions on the height of sunk costs a single company can bear per period, neither one of the possibilities seems plausible. Moreover, the reader should keep in mind that as the predating firm might be able to break free from her prisoner's dilemma, for all other firms the dilemma becomes even more severe. Thus, we obtain that in virtually all cases multi-market firms are faced with a prisoners' dilemma, which is at the heart of the excessive competition in the described economy. Having established this result, it is easy to understand why the inefficiency in the economy reduces alongside an increase in informational imperfection. The more costly it is for a firm to produce foreign goods, the less attractive it is for her to enter distant markets. Hence, at given sunk costs levels firms compete less, which means that the competitive and efficient transition-points approach each other.

6 Concluding Remarks

In this paper we presented a novel model of market entry in a multi-market economy. To this end, we employed a network perspective in order to investigate inter-market and inter-firm effects and proposed a framework that contributes to both the theoretical industrial organization literature as well as the literature on strategic network formation. In particular, for a multi-market economy of conglomerate firms, we obtained a complete characterization of efficient and competitive entry networks and showed that a general tension exists between them. Moreover, despite ex-ante symmetric assumptions, the model allows for asymmetric equilibria, granting extraordinary profits to firms with predatory entry strategies. However, we generally observe that firms are faced with a prisoners' dilemma, because individual and collective rationale point in opposite directions. This new observation was made possible because we did not follow the traditional practice of stipulating anonymous entrants. To the contrary, we explicitly allowed for a bilateral entry threat, which went on to manifest itself in the prisoners' dilemma. We also proved that the prisoners' dilemma and subsequent social welfare loss can be mitigated by a decrease in the exogenous informational spillover. This is a remarkable finding, since it promotes the use of informational barriers, such as patents, not only to secure innovation via the granting of temporal monopolies, but also as an effective instrument to control for welfare inefficiencies.

The framework established in this paper can be extended in several ways to study more aspects of interacting conglomerate firms in a multi-market environment. In particular, one should make a point of replacing the current ad-hoc equilibrium concept with a dynamic one in order to further reduce the set of plausible competitive entry networks. Another fundamental extension would be to allow for economies of scope and scale, which were not addressed in the investigation (although their importance is recognized). Moreover, when considering firms with a limited budget, especially predatory behavior would become infeasible, since the predating firms could not afford to invest large amounts of market entry costs. Furthermore, should we decide to shift the focus to heterogeneous firms, the model provides foundation for a multi-dimensional location model, where new firms could choose where to locate their core competencies. Finally, the technical framework may also be reinterpreted to model international trade and tariff games. In such a setting, firms would become countries and the informational spillover would reflect a tariff imposed on foreign goods.

References

Bala, V. and S. Goyal (2000): "A Noncooperative Model of Network Formation," *Econometrica*, 68(5), pp. 1181–1229.

Billand, P. and C. Bravard (2004): "Non Cooperative Networks in Multimarket Oligopolies," *International Journal of Industrial Organization*, 22(5), pp. 593–609.

Goyal, S. and S. Joshi (2003): "Networks of Collaboration in Oligopoly," *Games and Economic Behavior*, 43, pp. 57–85.

Goyal, S. and J. L. Moraga-Gonzalez (2001): "R&D Networks," *RAND Journal of Economics*, 32(4), pp. 686–707.

Neven, D. J. (1989): "Strategic entry deterrence: recent developments in the economics of industry," *Journal of economic surveys*, 3(3), pp. 213–233.

Wilson, R. (1992): "Strategic Models of Entry Deterrence," in: R. J. Aumann and S. Hart (eds.), *Handbook of Game Theory with Economic Applications*, vol. 1, chapter 10, Elsevier Science Publishers, North-Holland.

Part III

Agent Behavior and Market Outcome

Portfolio Selection in an Artificial Stock Market

Jörn Dermietzel[1], Detlef Seese[1], Thomas Stümpert[1], and Marliese Uhrig-Homburg[2]

[1] Institute of Applied Informatics and Formal Description Methods (AIFB),
Universität Karlsruhe (TH)
{dermietzel,seese,stuempert}@aifb.uni-karlsruhe.de

[2] Institute for Finance, Banking, and Insurance,
Universität Karlsruhe (TH)
derivate@fbv.uni-karlsruhe.de

Summary. *Agent-based computational finance* is a subfield in the rising field of *agent-based computational economics (ACE).* The idea behind this field of research is to validate theoretical models and to rebuild observed phenomena in a closed, completely controllable environment. Financial markets constitute a promising application for this field. Although financial theory has a broad foundation, its classical models are unable to explain the so-called *stylized facts* of real-world financial time series. Following the late 1980s a string of market models were presented that managed to reproduce these phenomena by simulating the interaction of heterogeneous traders. Most of them shared a common setup of one risky and one risk-free asset. In a recent survey of the field, LeBaron suggested abandoning this restrictive setup in order to enrich the observed dynamics (see LeBaron, 2006). He compares the transition from one to several risky securities with the replacement of one representative agent with heterogeneous agent models. Our goal is to extend the standard market setup to a multi-asset market model.

The main focus of this paper is on the portfolio selection of agents, which was not a major issue in the classical setup. Moreover, we present a new agent design that does not imprint fundamentalist or chartist behavior, but enables agents to evaluate different aspects of the market situation in order to maximize their expected utility. We implement a first type of myopic agents that use CRRA utility functions with heterogeneous degrees of risk aversion. These agents interact on an artificial stock market with stochastic dividends and stepwise price adjustment. We show that portfolio selection based on past dividends leads to converging prices that are in line with CAPM and thereby express the trade-off between risk and return. However many of the assumptions advanced in the theoretical model are not born out.

1 Introduction

In the later part of the 20th century researchers developed models of artificial stock markets (comp. LeBaron, 2000). The focus of these models was twofold: one aim was to understand market dynamics. A second aim was to understand the microscopic interaction of traders on the market, which produces observable macroscopic outcomes, which could then be compared to time series from real markets. Some of these models were able to explain phenomena observed on real markets yet not addressed by classical models in financial theory. These phenomena are today known as *stylized facts* (comp. Cont, 2001) and provide a benchmark for the outcomes of artificial stock markets.

One way to model financial markets is to interpret markets as interacting groups of learning, boundedly rational agents. This schema is called *agent-based computational finance.* LeBaron provides a good overview of this field and suggests questions for further research (see LeBaron, 2006). He also makes recommendations for tweaking parameters, such as increasing the number of assets in the market models. There have in fact been very few models with multi-asset settings up to now. The vast majority of these still limits the market to one risk-free and one risky asset. The few remaining models make simplifications at the trader level by assuming zero-intelligence traders (e.g. Cincotti et al., 2005) who use random allocation strategies, or limit traders to investing in only one risky asset at a time (e.g. Westerhoff, 2004).

This paper extends the standard setting to a multi-asset market in which traders manage portfolios with several positions. Due to the previously mentioned model design, portfolio selection was not a big issue in this area until recently. Traders on a multi-asset market need a working method for portfolio selection based on some kind of belief about market dynamics. This paper presents a new agent design, including advanced portfolio selection, and tests the outcome for consistency with financial theory.

The paper is organized as follows:
In the second section we present the market model and the general settings of the market. Afterwards we present the design of the trading agents. This design is tested on so-called *Dividend traders*, who are able to take into account the uncertainty of stochastic dividend processes but ignore the risk of price changes. The third section introduces some simulation results and compares these to the predictions of the *Capital Asset Pricing Model (CAPM).* The paper closes with a summary and some concluding remarks.

2 Model Description

We are looking at a stock market that is organized in time periods $t = 1, ..., T$. In each period artificial traders decide on their desired portfolio-composition and report their supply and demand to the market, i.e. the auctioneer. The latter then manages the trading process between agents, paying dividends and interests. Afterwards, he determines the new price for each stock.

There are I agents acting as traders on the market. Every trader can invest in a risk-free asset called a_1 and in $J-1$ risky assets $a_2, ..., a_J$. The risk-free asset, e.g. a bond, pays fixed interest $d_1 = r$, while the risky assets, e.g. stocks, pay stochastic dividends d_j, $j = 2, ..., J$.

2.1 Price Process

The traders are allowed to buy bonds from an imaginary bank and sell them at a fixed price of 1 per unit, but they are not allowed to sell short any asset or money. A constant number of shares of each stock is initially distributed equally among traders. Since the imaginary bank does not buy or sell stocks, traders can only sell stocks when there is sufficient demand, and they can only buy stocks if there is an adequate supply from other traders on the market. Within these restrictions the traders are allowed to change their portfolio each period. They therefore calculate their desired portfolio and report their demand $\delta^+_{i,t,j}$ and supply $\delta^-_{i,t,j}$ for each stock $j = 1, ..., J$ to an auctioneer.[1] The latter coordinates the trades between agents. If there is not enough supply to satisfy all demands for an asset j, i.e.

$$\sum_{i=1}^{I} |\delta^-_{i,t,j}| < \sum_{i=1}^{I} \delta^+_{i,t,j},$$

all demands for that asset are cut to

$$\delta^{*+}_{i,t,j} = \frac{\sum_{i=1}^{I} |\delta^-_{i,t,j}|}{\sum_{i=1}^{I} \delta^+_{i,t,j}} \delta^+_{i,t,j}.$$

Analogously, if the supply for an asset is not covered by demand, it is cut to

$$\delta^{*-}_{i,t,j} = \frac{\sum_{i=1}^{I} \delta^+_{i,t,j}}{\sum_{i=1}^{I} |\delta^-_{i,t,j}|} \delta^-_{i,t,j}.$$

The traders thus do not necessarily achieve their desired portfolios in every period.

After all possible trades have been executed, the auctioneer determines the new prices according to

$$p_{j,t+1} = p_{j,t} + \alpha \frac{\sum_{i=1}^{I} \delta_{i,t,j}}{p_{j,t}}, \qquad j = 2, .., J.$$

That means if there was excess demand for asset j in the actual period, the price is increased; otherwise, it is decreased. The scaling factor α manages the trade-off between fast reaction and stability of the market.

2.2 Information Representation

All relevant market variables are logged in a central matrix in which the rows and columns correspond to simulation periods and market-aspects, respectively. These aspects mainly consist of the prices $p_{t,1}, ..., p_{t,J}$ and dividends $d_{t,1}, ..., d_{t,J}$. It is possible to extend the market matrix by arbitrary aspects characterizing the status of the market, e.g. individual trader attributes, such as demands or portfolios, or relevant environmental information. Since the market matrix is accessible to all agents, the information included can be interpreted as *public information.* It is therefore an important design issue which aspects should be included in the market matrix, and which should be kept private, i.e. recorded in a decentralized manner. In this paper the market matrix is restricted to the prices and dividends of the assets.

The status $s_t \in S = \mathcal{R}^Z$ of the market is characterized by the corresponding row t of the market matrix, that is $\{m_1, m_2, ..., m_Z\}$.[2] All past market states can be arranged in

[1] Of course the agents do only report one value $\delta_{i,t,j}$ for each asset to the auctioneer. The flags $-$ and $+$ indicate if it is supply, i.e. $\delta_{i,t,j} < 0$, or demand, i.e. $\delta_{i,t,j} > 0$. The distinction between demand and supply, i.e. negative demand, is only done when needed otherwise $\delta_{i,t,j}$ is used to capture both.

[2] $S = \mathcal{R}^Z$ is the most general market space. Depending on the aspects represented in the market matrix, one can restrict single dimensions, e.g. to $\mathcal{R}_+$ or $\mathcal{N}$. The length Z of the market vector depends on the number of market-aspects to be represented. In this paper the market vector contains the prices and dividends, which leads to $Z = 2 * J$.

a transition graph in which each node stands for a certain market status s. The edges, e.g. $\overline{ss}'$, are weighed with the number of observed direct transitions from status s to s', i.e. the weight of a transition at time t is given by $\overline{ss}' = f_t(s, s') : S \times S \rightarrow \mathcal{N}_0^+$. From this graph, one can derive the historical probabilities of moving to a certain market status s' given the actual status s. That is

$$P_t(s_{t+1} = s' \mid s_t = s) = \frac{f_t(s, s')}{\sum_{s^*=S} f_t(s, s^*)},$$

where s^* are all possible market states to follow s.

Unlike in most other simulation models, the quantities of stocks, such as supply and demand, are not expressed in shares of stocks but in monetary units (MU). The agents determine the composition of their portfolios, i.e. every agent determines the amount of money he wants to invest in each asset at given prices. That is analogous to theoretical portfolio selection in finance, where each investment is normed into fractions of wealth (see Markowitz, 1952). One can easily convert the monetary units into shares, but it is easier from the logical point of view (and more consistent) to deal with monetary units throughout the whole model.

2.3 Forecast of Market State s_{t+1}

All traders have access to the market matrix. As described above, one can derive empirical probability distributions to forecast the next market state from this matrix. This only works if every possible market state has occurred often enough to derive a smooth probability function, which is unlikely for complex market vectors. Therefore, agents use individual views, which reduce the complexity of the market vector.

First, each trader has to form tuples of market-aspects that correlate to his belief with respect to the market dynamics. These tuples are disjoint subvectors of the market vector. For example, let the market vector be represented by $\{w, x, y, z\}$: if market-aspect w and x are correlated, while there is no such connection to or between y and z, then the trader may form the tuples $\{w, x\}, \{y\}$. We call these tuples of correlated market-aspects *groups*. If an aspect is not represented in any of these groups, like aspect z in the example, the trader will not expect this one to change. To forecast the aspects within each group, the trader defines a filter for each group. The filter shows up on which market-aspect's past realizations the expected realizations of the particular group are based, and what the relation between these is. We therefore call the result of the filtering process the *basis* of a certain group. That is the pattern the trader is searching for in order to forecast the future realizations of the aspects of the corresponding group. A filter for a group is a tuple $\{m_1, m_2, ..., m_Z\}$ with $m \in \{\mathcal{N}_0^+, \star, \#\}$. Numbers and the $\star$ symbol are referred to as *matching symbols* while the # sign is called an *ignoring symbol.* The numbers stand for the length of the pattern the trader is searching for, i.e. 0 means that the trader is searching for the status s_t, 1 indicates that the trader is searching for a pattern $\{s_t, s_{t-1}\}$ to have occurred, and so on. The $\star$ symbol matches every value for that aspect, i.e. the agent expects this aspect's realizations to be independent from past realizations. The # symbol causes an aspect to be ignored. Given a group and the corresponding filter, the trader can simulate the realization of the aspects in the group by drawing a sample of size K from the states that followed the pattern he was searching for. After that, he holds K tuples of values for the aspects of the group of market-aspects, which reflect the transition-probabilities given above. The forecasting is finished by combining the value-tuples to K complete market status vectors $\{w, x, y, z\}$.

2.4 Portfolio Optimization

The simulation of market states is based on the history of the market. The results are K market states $\hat{s}_k$; these reflect the historical transition-probabilities of the market and the individual beliefs of the traders with respect to market dynamics. One can easily calculate the return of each asset given the actual market status s_t and the K forecasted market states $\hat{s}^k_{t+1}$, $k = 1, ..., K$:

$$\hat{r}^k_{j,t+1} = \frac{\hat{p}^k_{j,t+1}}{p_{j,t}} + \frac{\hat{d}^k_{j,t+1}}{p_{j,t}} - 1, \qquad j = 1, ..., J.$$

Now the trader can optimize his desired portfolio $\gamma_{1,t+1}, ..., \gamma_{J,t+1}$ with respect to the forecasted returns in order to maximize the utility of his one-period-ahead gain. That is

$$\max_{\gamma_{1,t+1},...,\gamma_{J,t+1}} \frac{1}{K} \sum_{k=1}^{K} u \left(\sum_{j=1}^{J} \gamma_{j,t+1} * \hat{r}^k_{j,t+1} \right), \qquad s.t. \sum_{j=1}^{J} \gamma_{j,t+1} = w_t,$$

where w_t is the actual wealth of the trader, i.e. his budget, which starts with an exogenous endowment in $t = 1$ and evolves with

$$w_t = \sum_{j=1}^{J} \left(\gamma_{j,t} \frac{p_{j,t} + d_{j,t}}{p_{j,t-1}} \right).$$

Since this maximization problem is not easily solved the analytic way, and we do not want to restrict the model to a special family of utility functions, the trader uses the following heuristic to optimize his portfolio:
Start with the actual portfolio $\gamma_{1,t}, ..., \gamma_{J,t}$ as the desired portfolio.
Repeat the following steps L times:

1. Pick two assets j and j' at random.
2. Transfer money from j to j' until this no longer increases the expected utility, or there is no positive amount of money left in asset j.
3. The resulting portfolio $\gamma^*_1, ..., \gamma^*_J$ is strictly preferred to the desired portfolio, and thus replaces the latter.

The simulation results in the next section will show how well this heuristic approximates the optimal portfolios.

3 Validation of Portfolio Decision Based on Dividends

The first types of agents to analyze are pure dividend traders. This type of agent does not take price movements into account. Given the market vector $\{p_{1,t}, p_{2,t}, p_{3,t}, r, d_{2,t}, d_{3,t}\}$ for a market with two risky assets, the dividend traders use the groups $\{d_1\}, \{d_2\}$ and $\{d_3\}$ with the filters $\{\#, \#, \#, \star, \#, \#\}$, $\{\#, \#, \#, \#, \star, \#\}$ and $\{\#, \#, \#, \#, \#, \star\}$.

This means they assume the prices will not change and that the dividends will be independent. Because the dynamics are restricted to the impact of the exogenous given dividend-processes, this setup allows us to compare the simulation results to the predictions of the Capital Asset Pricing Model (*CAPM*).

We simulate 10 runs of $T = 5000$ periods each, where the first 10 periods are neutralized, i.e. there is no trade and the prices are kept constant to provide a minimal number of dividends upon which to base the forecasting. There are only 9 agents, $i = 1, ..., 9$, acting on the market. These traders are dividend traders and use utility functions with constant relative risk aversion (*CRRA*) of the form:

$$u(w) = \frac{1}{\delta} w^{1-\delta},$$

with $\delta = i/10$, which means the risk aversion corresponds to the ranking of the traders. Agent 1 is the least risk-averse trader, while trader 9 is the most risk-averse. There are two risky assets in the market, which pay stochastic dividends $d_{j,t} = \mathcal{N}(\mu = r * p_{j,0}, \sigma_j^2)$, with $\sigma_2 = 16$, $\sigma_3 = 25$ as standard deviations and $p_{j,0} = 1000$ as the price for one share of the risky assets at time $t = 0$, which are all given exogenously. The initial distribution of shares is also given exogenously, i.e. every agent is endowed with a portfolio $\gamma_{i,j,0} = \frac{2000}{J}$. The sample size K is set to 500 and the number of iterations is $L = 15$.

The predictions of CAPM are based on important assumptions, some of which are not met in our model:

1. Traders act as price-takers, i.e. they behave as if their individual decision does not affect the prices.
2. Traders can sell short assets without limits, which includes borrowing unlimited amounts of money at risk-free interest.
3. The market is assumed to be in equilibrium, i.e. at the price at which no trader wants to change his current portfolio. In other words, the market is cleared, and excess demand and excess supply are both at zero.

The last assumption is especially important for determining the optimal portfolios for traders analytically, which leads to the fundamental results of the CAPM. Our model does not start at equilibrium, and the traders are not allowed to sell assets short, including the risk-free asset a_1. Although these assumptions are not met, our simulation-market is expected to converge to a status that approximates the equilibrium status of the theoretical model, which would prove that the forecasting algorithm and the portfolio optimization do work.

The simulation setup is designed to validate the portfolio decisions made by agents. We therefore compare the final prices for $\sigma_2 = 16$ with those for $\sigma_3 = 25$. The price for the riskier asset is expected to be significantly lower than that of the asset with lower risk. In addition, we compare the fraction of wealth invested in the risky assets across agents. This fraction is expected to correspond to the individual degree of risk-aversion. We show that the simulation results support important theoretical results, such as the Tobin-Separation and the market line.

Figure 1 illustrates a typical simulation run with two risky assets. One can clearly see that the price reflects the expected dividends for both assets and thereby keeps the expected return at an almost constant level. With increasing periods the expected

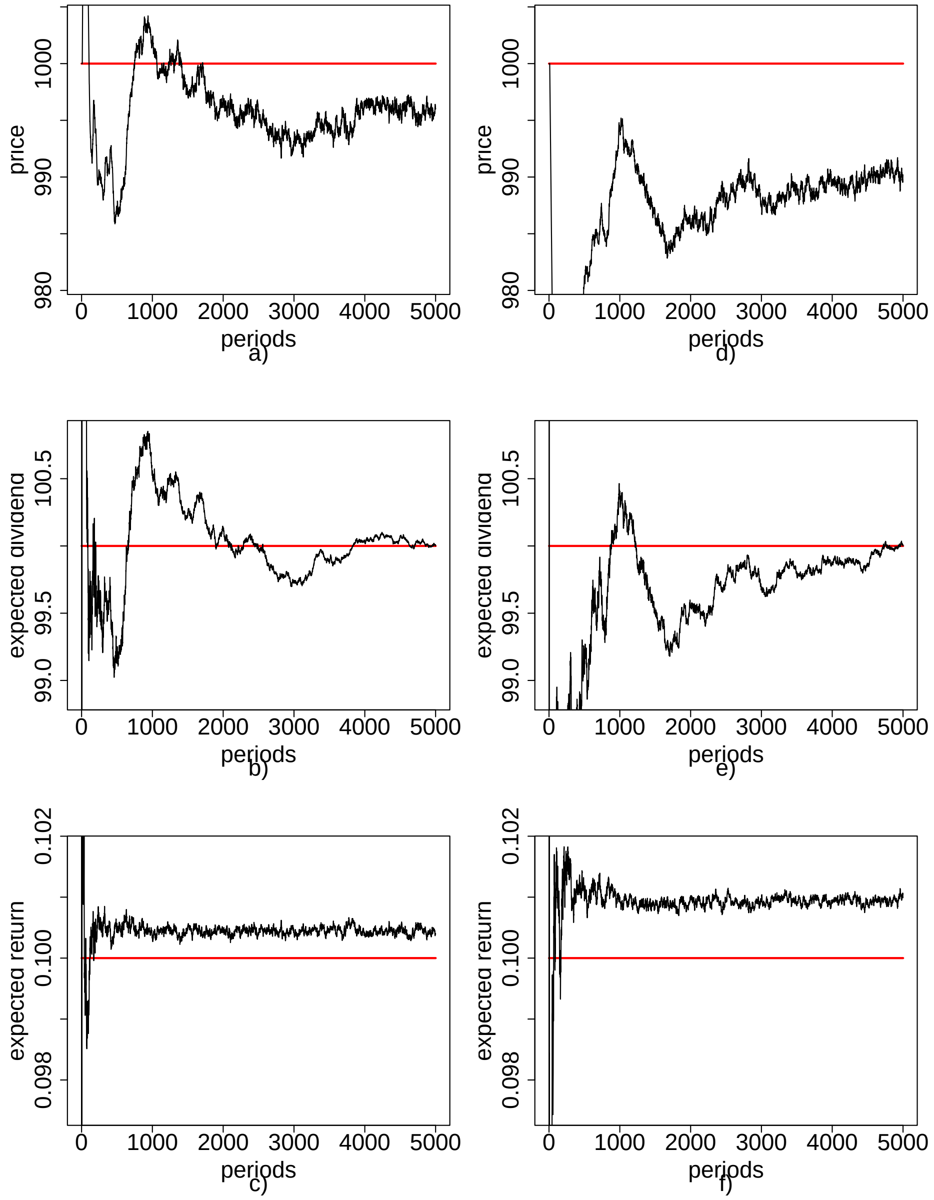


Figure 1: Typical simulation output for a simulation run with two risky assets. Left column (a-c): asset with $\sigma = 16$; right column (d-f): asset with $\sigma = 25$.

dividend approaches $r * p_{j,0} = 100$, and the price stabilizes at a level which is clearly below the risk-neutral price of 1000 MU for both risky assets. To compare the final prices, we calculated the average prices for each stock over the last 10 periods for each run. With 995.5375 MU for a_2 with $\sigma_2 = 16$ and 990.8640 MU for a_3 with $\sigma_3 = 25$, the difference of the average final prices over all 10 runs is clearly significant.[3]

Another indicator for the proper working of the portfolio selection is the proportion of wealth invested in the risky asset. Since the agents are heterogeneous in their degree of risk aversion, their desired portfolios should also differ systematically. Figure

[3] Significance (p-value $>$ 0.9998) is tested with the Mann-Whitney-Test, which tests if the two samples are taken from identical populations, resp. if their means differ.

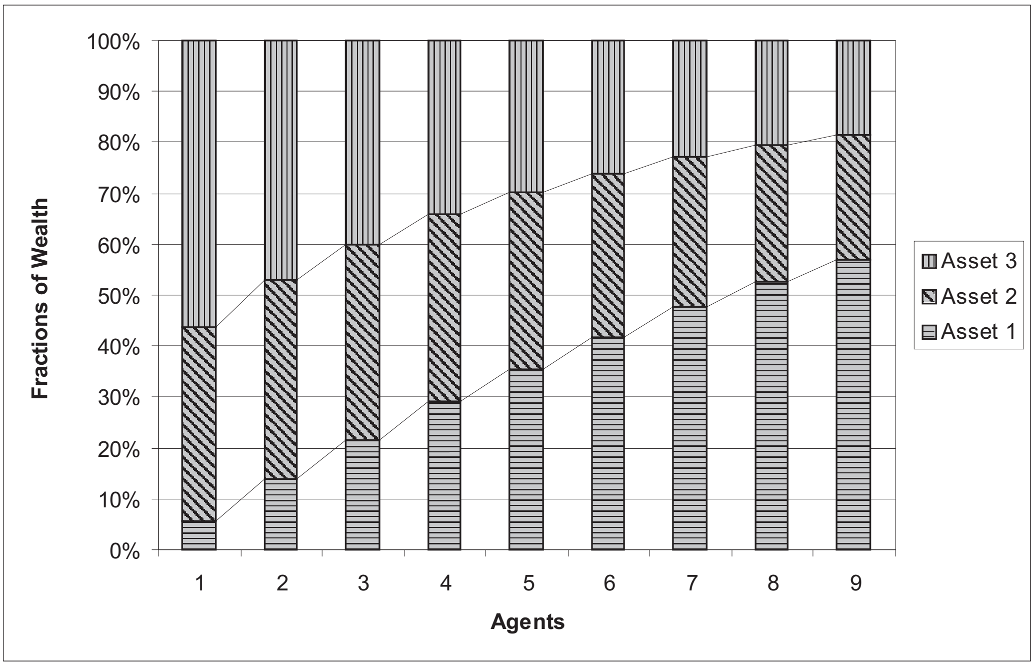


Figure 2: Average portfolio composition, i.e. fractions of wealth invested in a_1 to a_3, for the least risk-averse agent $i = 1$ to the most risk-averse agent $i = 9$.

2 illustrates the proportions of wealth invested in the different assets for all agents averaged over all periods. One can clearly see that the agents' degree of risk aversion is reflected in their portfolio composition. The fraction of money invested in the risk-free asset a_1 significantly increases from the least risk-averse agent $i = 1$ to the most risk-averse agent $i = 9$.

Figure 3 shows the μ-σ-plot for the average simulation outcomes, i.e. the average expected returns from the dividends in the last 10 periods, which are $\mu_2 = 0.10045$ and $\mu_3 = 0.10092$, as well as the corresponding risk $\sigma_2 = 0.01617$ and $\sigma_3 = 0.02479$. From these data one can derive the efficient border and the market line, which touches the efficient border in the market portfolio. As one can see in Figure 3, all portfolios are on the market line with the exception of the one belonging to the least risk-averse agent. The latter desires a portfolio on the market line, where more than 100 percent of his wealth is invested in the market portfolio. In other words, he would like to sell the risk free asset $a_1$short and invest this additional money in the market portfolio. Because this is not allowed in our model, the agent has to realize a portfolio off the market line by selling a_2 in order to invest more in the riskiest asset a_3.

4 Conclusion

We presented a new algorithm to forecast a multi-asset market as well as a heuristic to optimize portfolios based on these forecasts. Furthermore, we implemented a complete simulation market to test these algorithms in a closed and controllable environment.

The simulation runs showed that the market outcome clearly reflects the level of risk aversion among the agents. The interaction of the agents led to prices that stabilized the expected returns from dividends at a level above the risk-free return, which is a

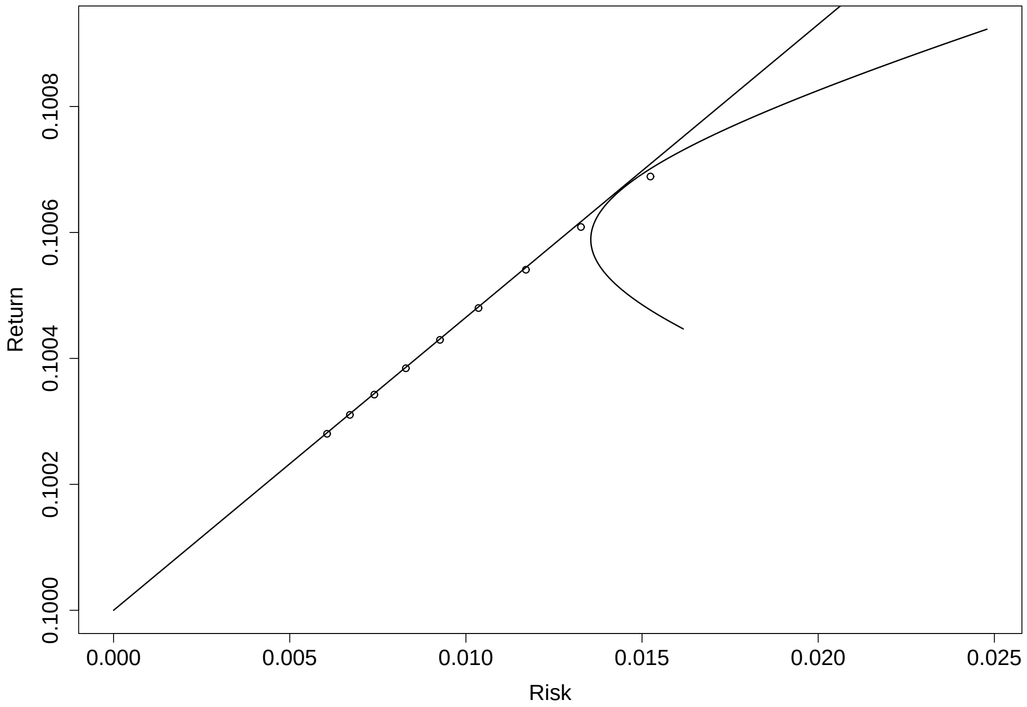


Figure 3: μ-σ-Plot for the average outcomes over the 10 simulation runs: efficient border, market line and average agent portfolios

clear message in terms of risk compensation. The average agent portfolio composition also reflects the individual degree of risk aversion.

The comparizon to the theoretical predictions of the corresponding CAPM shows how well the algorithms for forecasting and portfolio optimization work. Although the model started from an off-equilibrium point, and some assumptions from the theoretical model were not met, the interaction of the traders led to a fast and stable approximation of the theoretical outcome.

The next step is to invent and implement a learning algorithm that allows agents to adapt their individual beliefs to the market dynamics. It will be of great interest whether the market will then continue to approximate the theoretical outcome or if it will produce more complicated dynamics that do not necessarily converge.

References

Cincotti, S., L. Ponta, and M. Raberto (2005): “A multi-assets artificial stock market with zero intelligence traders,” working paper, DIBE-CINEF, Universita di Genova.

Cont, R. (2001): “Empirical properties of asset returns: stylized facts and statistical issues,” *Quantitative Finance*, 1(2), pp. 223–236.

LeBaron, B. (2000): “Agent-based computational finance: suggested readings and early research,” *Journal of Economic Dynamics & Control*, 24, pp. 679–702.

LeBaron, B. (2006): *Tesfatsion, Leigh S. and Judd, Kenneth L.: Handbook of Computational Economics, Vol. 2: Agent-Based Computational Economics*, chapter Agent-based computational Finance, North-Holland, Amsterdam.

Markowitz, H. (1952): “Portfolio Selection,” *Journal of Finance*, 7, pp. 77–91.

Westerhoff, F. (2004): “Multiasset Market Dynamics,” *Macroeconomic Dynamics*, 8(5), pp. 596–616.

Utilizing Financial Models in Market Design: The Search for a Benchmark Model

Yalın Gündüz[1], Marliese Uhrig-Homburg[1], and Detlef Seese[2]

[1] Institute for Finance, Banking, and Insurance, Universität Karlsruhe (TH)
{yalin.gunduz,derivate}@fbv.uni-karlsruhe.de
[2] Institute of Applied Informatics and Formal Description Methods,
Universität Karlsruhe (TH)
seese@aifb.uni-karlsruhe.de

Summary. One key element of successful market engineering is validating the design through a benchmark financial model. Although there are competing frameworks in the field of credit risk, there has not been any agreement on which approach better fits the observed prices. For this purpose, we empirically compare the out-of-sample prediction errors of basic forms of structural and reduced-form models, in a cross-sectional setup. Moreover, results are contrasted to the cross-sectional performance of a machine learning algorithm. The results shed light upon the steps towards a benchmark credit risk model that can be utilized in the market design process.

1 Introduction

Successful electronic market design necessitates the careful integration of many aspects: The market microstructure, the infrastructure and the business structure all have to be determined and incorporated into the design process (Weinhardt et al., 2003). One important step in this process is the validation of the design. Which evaluative tool can the designers rely upon to test their model? During the validation phase, different market settings can be simulated with the help of a benchmark financial model to assure the robustness of the design with respect to financial expectations. Having developed rapidly in the last decades, financial modeling practice has commonly sought to reach a fair price for different instruments. The models that best fit the observed prices have not only provided an accurate estimate for the fair price, but have also been used as a benchmark for market participants in their actions. In equity and FX derivatives, applying the Black/Scholes option pricing framework has been widely accepted as the benchmark model. This price may be used to examine hedge ratios and arbitrage possibilities, and is treated as fundamental information. Black/Scholes prices compare well to actual prices in the market and provide guidance in market actions. However, there has been no single pricing model that would serve as a benchmark in the area of credit risk, the field that investigates reaching a fair price for instruments carrying risk of default. There are several approaches to modeling this risk. On the one hand, there are structural models that are based on modeling of the evolution of the issuer's balance sheet. On the other hand, reduced-form models specify the credit risk exogenously. The efforts to establish a widely accepted framework are ongoing, with plenty of room for further development. Theoretical considerations aside, the current stalemate is rooted in the fact that previous empirical studies have yielded controversial results. It is therefore necessary to re-evaluate what the frameworks offer, and empirically test the two credit risk modeling frameworks in parallel, which has not yet been investigated in the literature. For this purpose, credit default swaps, which are highly liquid instruments, are ideal. This study empirically compares the out-of-sample prediction quality of the basic structures of structural (the Merton model) and reduced-form models (constant intensity) in a cross-section of firms. In addition to the test results of the financial models, a third method's prediction results are presented in this study. The Support Vector Machines (SVM) algorithm is a new machine learning technique for data regression. Having opened a new pathway for computing empirical predictions, SVM, whose fundamentals were developed by Vapnik (1995), has become a competitive alternative to traditional neural network approaches due to its empirical performance. Overall, as our empirical efforts point a benchmark model, the results can be used as an indispensable validation tool in designing markets of credit-risky instruments.

The rest of the study is organized as follows: Section 2 describes the credit default swap market and the interdealer brokerage data used in the study. Section 3 provides an overview of the literature on financial credit risk models and SVM algorithms as well as the pricing structures for credit default swaps. Section 4 presents the out-of-sample prediction results with financial and SVM methods. In the last section we will make concluding remarks and suggest issues for further research.

2 Credit Default Swap Market and Dataset

With 51 per cent of the market share in a rapidly expanding market, credit default swaps are by far the most frequently traded type of credit derivatives (British Bankers' Association, 2004, p. 21). A credit default swap (CDS) is an insurance contract against the default of a specified commercial or sovereign entity. By entering into a CDS contract, the buyer of the CDS makes periodic premium payments to the seller of the CDS, who is obliged to compensate the buyer in case of default of the underlying entity. This premium is expected to increase in the entity's default risk; thus, CDS prices are a nearly ideal measure of credit risk. Moreover, being contracts rather than securities, CDSs are less constrained by supply and demand pressures. This liquidity makes them an obvious choice for empirical analysis.

To illustrate the mechanics of a CDS, let us suppose, e.g. on June 20, 2005, that an insurance buyer agrees to enter into a 5-year CDS contract with a seller, written on a bond issued by DaimlerChrysler AG, with a CDS premium of 80 basis points, on a contract notional amount of USD 5 Million. The buyer may or may not own the corporate bonds issued by DaimlerChrysler. Let us assume the buyer owns 5,000 of the underlying corporate bonds that mature on April 15, 2010, each having a par value of USD 1,000, so that the buyer would have fully covered protection against any potential loss $(5,000 \times 1,000)$. In exchange for the protection, the buyer has to pay quarterly installments of approximately $1/4 \times 80$ basis points of the notional amount (depending on the actual days in a quarter). In monetary terms, this corresponds to quarterly payments of $5,000,000 \times 1/4 \times 0.0080 =$ USD $10,000$. The buyer will be paying this quarterly amount for 5 years. If default does not occur, the seller pays nothing. If it occurs during the lifetime of the CDS, there are two forms of settlement: In a physical settlement, the buyer delivers the eligible bonds of the defaulted entity to the seller, in exchange for the contract notional amount. The alternative to this is a cash settlement, in which the buyer keeps the underlying bonds, but is compensated for the loss. In both instances, the buyer's loss is fully covered by the notional amount.

For the empirical study, the CDS datasets were retrieved from CreditTrade's daily indicative bid/ask quotes. While credit default swaps are traded OTC, the emerging role of interdealer brokers (IDBs) is worth noting. As an alternative to direct phone trading, dealers from major institutions can post quotes to IDBs through either voice brokers or electronic brokerage platforms. Figures for 2004 show that broker-intermediated trades constituted 34 per cent of the total in the credit derivatives market (International Swaps and Derivatives Association, 2004). The interesting highlight from the credit default swap market is the gradual shift from voice broking to electronic platforms. This phenomenon is consistent with the history of today's more mature markets. Many IDBs have initiated electronic platforms, but have nevertheless opted to keep the voice broking alive, which suggests its importance in the value chain. "The market color" that the voice broker provides is still seen as an important ingredient when complex, illiquid, and larger size transactions come into the picture. Gündüz et al. (2006) investigate the microstructural issues in the credit default swap market, providing evidence from different market structures.

The liquidity of the CDS market has increased over time. This is reflected in the bid-ask spreads, which can be taken as a proxy for the liquidity (see Schwartz and Francioni (2004) for a discussion on liquidity). Table 1 presents the steadily decreasing bid-ask spreads over time. CDS of the entities that are denominated in US Dollars tend to have a wider bid-ask spread across the full horizon of December 2002-January 2005.

Table 1: Number of Observations and Bid-Ask Spreads for the Full Sample

		2002		**2003**		**2004**		**Total**	
Type	Currency	Obs	Spread	Obs	Spread	Obs	Spread	Obs	Spread
Corporate / Bank	EUR	3420	20.51	43152	11.85	45230	5.57	91802	9.08
Corporate / Bank	USD	1558	32.56	19841	24.06	20066	15.76	41465	20.36

Obs: Number of Observations in the Cluster
Spread (bps): Average of the Difference of Offer Price - Bid Price

A cross-sectional design necessitates the formation of risk classes that have the same risk properties. This would indicate that the companies in these risk classes are attributed to price credit risk similarly. The most obvious division is due to ratings. Although there are both investment grade and high-risk CDSs in the dataset, the liquid segments of Aa and A risk classes are focused on. Moreover, 80 per cent of the data tabulated in Table 1 are CDSs with 5-year maturity; they are thus an obvious selection. Another breakdown is according to the rank of the instrument. CDSs written on senior underlyings have a priority in payment in comparison to subordinate underlyings. This should also be reflected in CDS premiums, as senior CDSs contain less risk premium. Our data has been split up accordingly. One final breakdown is according to region and currency. In the dataset, we had two regions, where all entities in a region were denominated in a single currency. Therefore, the dataset is further divided into two risk classes, European (Euro-denominated) and North American (US Dollar-denominated). Following the construction of six risk clusters, the remaining dataset had over 55,000 price observations. As can be observed in Table 2, on average, North American CDSs have higher premiums and wider bid-ask spreads. The observation of the increase of average premiums with decreasing credit quality in all three moves from Aa to A is in line with the theoretical argument that the higher the risk of default, the higher the fee should be to obtain insurance.

Table 2: Risk Classes and Their Average Midpoints, Spreads, and Number of Observations

	Europe, EUR			**N.America, USD**			**Total**		
	Mid	Spread	Obs	Mid	Spread	Obs	Mid	Spread	Obs
Aa Senior	18.63	4.75	10901	30.85	10.37	5046	22.50	6.53	15947
Aa Subordinate	30.17	6.08	3301	-	-	-	30.17	6.08	3301
A Senior	42.49	6.81	20212	46.53	12.69	11139	43.93	8.90	31351
A Subordinate	41.88	7.09	4465	-	-	-	41.88	7.09	4465

Mid (bps): Average of the Midpoint of each Bid and Offer Price
Spread (bps): Average of the Difference of Offer Price - Bid Price
Obs: Number of Observations in the Cluster

3 Pricing of Credit Default Swaps

3.1 Structural Models

Central to the development of the financial engineering field, the seminal work of Black and Scholes (1973) presented a closed form solution to price stock options. Shortly thereafter, the study of Merton (1974) viewed equity as a call option on firm's assets, which enabled him to make use of the Black and Scholes framework. Consequently, the study has reached a "risk structure" of zero-coupon bonds. At the bond's maturity, if the asset value that follows a diffusion process is above the debt value, the firm remains alive; otherwise, default occurs. Due to making use of firm-specific financial parameters, this approach has been named as "structural" modeling. Merton's model has been extended by allowing default before maturity (Black and Cox, 1976), enabling default in case of unwillingness to meet obligations and endogenous determination of the default boundary (Leland, 1994; Leland and Toft, 1996; Anderson and Sundaresan, 1996; Goldstein et al., 2001), and stochastic interest rates (Longstaff and Schwartz, 1995). On average, however, empirical studies with structural models do not indicate a good fit. This was especially attributed to the estimation difficulties of the firm value and asset volatility parameters (Ericsson and Reneby, 2004, 2005). Early studies found that the Merton model consistently underpredicted spreads (Jones et al., 1984; Ogden, 1987; Lyden and Saraniti, 2000). Recent studies have extended this argument by pointing out that not all structural models act the same (Eom et al., 2004).

Most empirical studies have relied on bond prices as their source. However, as discussed in the previous section, the rapidly expanding markets of credit default swaps provide a new alternative for empirical analysis. It is very important to utilize highly liquid instruments in an empirical study. As the CDS price is comprised solely of the default premium, it can be regarded as a pure measure of credit risk. The possibility of adapting credit risk pricing models to their valuation makes CDS an important source for empirical studies in the area. With the exception of a few studies in the literature (Ericsson et al., 2005), CDS data has not been utilized for testing structural models. Our study extends available empirical studies as relying on CDS data as the single source for estimation and out-of-sample prediction.

For the purpose of comparability in our study, the most basic structures of each approach have been selected, with the Merton model being the representative for structural models. Pricing a CDS can be undertaken by valuing the premium leg, the insurance fee that the buyer has to pay to the seller of the CDS against default; and by computing the protection leg, the single compensation that the seller is obliged to pay to the buyer in case of default of the underlying. The Merton model can be adapted to CDS valuation by allowing default only at maturity. Accordingly, the premium leg's present value ($Prm(0)$) is the sum of discounted premiums paid until maturity (Equation (1)). The quarterly paid premiums start at 0.25, and end in 5 years, which is the maturity of the CDSs that are present in the dataset.

$$Prm(0) = p(theo) \sum_{i=0.25}^{5} e^{-r(i).i} = p(theo)D(T) \tag{1}$$

where $p(theo)$ is the theoretically fair premium and $r(i)$ is the riskless interest rate on day 0 for maturity i . To set up the protection leg's present value ($Prt(0)$), the non-recoverable portion is discounted from the maturity of the contract multiplied by the probability of default.

$$Prt(0) = (1 - \Psi)\Phi(A)(e^{-r(5)5}) = (1 - \Psi)C(T) \tag{2}$$

where $\Phi(A)$ is the risk-neutral default probability in the Merton setting, and A stands for the well-known Black and Scholes parameter. Ψ , the recovery rate, is the amount recovered in default, and it is taken as a constant. Finally, the no-arbitrage rule requires the premium and protection leg's value at contract initiation to be equal, which leads to the theoretically fair CDS premium that can be used as a benchmark.

$$p(theo) = \frac{(1 - \Psi)C(T)}{D(T)} \tag{3}$$

3.2 Reduced-Form Models

A second approach that has been popular since the last decade is modeling default probabilities as an exogenous variable represented by a default intensity. While remaining silent about the cause of default, the unpredictable default is modeled by a jump process. This constitutes the main difference between structural and reduced-form models. No diffusing firm value is assumed in the reduced-form setting; in contrast, the default intensity is the parameter for a Poisson process. Extensions of the constant intensity approach by Jarrow and Turnbull (1995) have altered this work by rating dependence of the intensities (Jarrow et al., 1997), or by placing the dependence of the intensities on one or more state variables (Lando, 1998; Duffie and Singleton, 1999). Empirical studies have also arisen in the area. Those conducted by Duffee

(1999), Bakshi et al. (2004) and Longstaff et al. (2005) are among the most important contributions. Uhrig-Homburg (2002) provides a comprehensive overview on both structural and reduced-form modeling.

So as to be comparable to the selection of the most basic structural model (Merton) employed in this study, the constant default intensity approach was selected. The simplifying aspects of the model enable a direct comparison with its structural counterpart. Unlike the Merton model, a constant intensity model allows default any time during maturity. In the intensity setting, the theoretically fair CDS premium is reached similarly by equating the premium and protection leg at contract initiation:

$$p(theo) = \frac{(1-\Psi)\sum_{i=0.25}^{5} e^{-r(i)i}(e^{-\lambda(i-0.25)} - e^{-\lambda.i})}{\sum_{i=0.25}^{5} e^{-(\lambda+r(i))i}} \tag{4a}$$

where λ is the default intensity parameter. However, with constant intervals between premiums, this reduces to Equation (4b), where the interest rate parameters are excluded in the final form.

$$p(theo) = (1-\Psi)(e^{\lambda\Delta i} - 1) \tag{4b}$$

3.3 SVM Regression

The third approach under analysis is the SVM Regression Method, where no financial structure is inherent in the algorithm. After the introduction of its fundamentals by Vapnik (1995), SVM methods have been competitive to neural network algorithms, and they have been shown to yield better results in many applications. The SVM Regression method uses special kernel functions to map the data space into a high dimensional feature space. This non-linear mapping is combined with a linear machine to regress the data, while mathematical properties of the kernel functions ease the computation. The key concept at work here is that the inner product in feature space has a corresponding kernel in the input space, so that simply taking the inner products of test and training variables is adequate (Cristianini and Shawe-Taylor, 2000). Although there are some financial time series forecasting applications (Cao and Tay, 2001; Müller et al., 1997), empirical literature on SVM Regression methods is still developing. The cross-sectional prediction capability of the SVM method has not yet been tested with CDS prices and its results have not yet been contrasted to financial methods.

We have selected four basic kernel types: linear, polynomial, Gaussian radial basis (RBF), and exponential radial basis kernel functions (ERBF). The selection of these basic kernels will enable a better comparison with structural and reduced-form models, which have also been selected in their simplest forms. The first case uses a linear kernel function with the inner products of test and training points, m and n, respectively (Equation (5)). Similarly, the following kernel functions in Equations (6)-(8) present the polynomial, Gaussian radial basis, and exponential radial basis functions in order. The polynomial kernel is a popular method for non-linear modeling, whereas radial basis functions have been greatly emphasized in recent studies.

$$K(m,n) = \langle m,n \rangle \tag{5}$$

$$K(m,n) = (\langle m,n \rangle + 1)^2 \tag{6}$$

$$K(m,n) = exp\left(-\frac{||m-n||^2}{2\sigma^2}\right) \tag{7}$$

$$K(m,n) = exp\left(-\frac{||m-n||}{2\sigma^2}\right) \tag{8}$$

where σ is taken to be 0.5. In all four cases, the cost function is an important parameter to select, which allows the slack in the system and works as a penalty parameter of the error term. This is also taken as default, 10, in all runs. Moreover, an ε -insensitive band, that has a value of 10E-4 is constructed in all SVM Regressions. These parameter values are selected as a default value, in accordance with the literature.

4 Empirical Analysis

4.1 Design

In order to test the out-of-sample prediction quality of financial and machine learning methods, companies in each of the six risk classes are divided into an estimation and a prediction sample. For financial models, default probabilities/intensities are estimated from Equations (3) and (4b), from the observed CDS premiums for each firm in the estimation sample for each day. So as to estimate these model parameters, the interest rates and recovery rates are needed. Riskless interest rate data is required for the Merton model. US Constant Maturity prices are used for North American CDSs, whereas the German Federal Bank's estimates for the Svensson (1994) parameters are utilized for the European region. Although the recovery rate could, in principal, be estimated jointly with the default probabilities/intensities, it has been taken as 0.5 based on the work of Altman and Kishore (1996), and following recent research and practice that have demonstrated the insensitivity of the results to selection (Frühwirth and Sögner, 2006; Houweling and Vorst, 2005).

After this step, each company's Black/Scholes parameter A and default intensity λ is averaged to obtain "aggregate" daily $\Phi(A)$ and λ values for each risk class. These values are plugged into Equations (3) and (4b) for a theoretical CDS premium on that day, which are then used to test out-of-sample prediction performance (i.e. whether the models have the ability to reach the observed CDS premiums of the companies in the prediction sample). Separation of the companies into samples has been on a random basis, and had a ratio of 2:1-3:1 with respect to estimation and prediction samples.

In a similar manner, SVM algorithms are tested in an out-of-sample design. However, the companies in the estimation sample that are used in the financial models have to be further divided into three groups. This stems from the fact that the CDS data of the first two groups of companies have to train the SVM function with an input-output mapping. The third set of firms in the estimation sample could then be used to predict the CDS premiums of the fourth, prediction sample. In each of the four samples the CDS prices on a given day for all firms are averaged in order to obtain a daily observation, in the spirit of the method we used with financial models.

4.2 Results

Table 3 presents the out-of-sample results for each of the six approaches, the two financial models and the four SVM kernels chosen. For each risk cluster, prediction errors and the number of observations in estimation and prediction samples are given. Prediction errors vary significantly across approaches. Whereas the financial models perform quite similarly within risk clusters, the performance of the SVM kernels varies substantially according to the kernel selected. In some risk clusters the polynomial kernel failed completely. The Gaussian radial basis (RBF) and the exponential radial basis (ERBF) kernel functions also clearly perform worse than the financial models. In the following comparison, we therefore concentrate on the linear SVM kernel which appears to have the best results within the SVM approaches tested.

Table 4 provides significance tests for the difference of the absolute errors for Merton, intensity, and linear kernel SVM. Results are reached with the Yule-Walker method, by means of the backward elimination of insignificant autocorrelation lags. At first glance, the prediction errors of the Merton model and the constant intensity model look similar, despite the structural difference of the Merton model allowing default at maturity. However, the significance tests provided in Table 4 signify that in three out of six risk classes, the absolute errors of out-of-sample prediction are lower for the Merton model, which has a more restrictive setup. One interpretation of this outcome could be that estimating the default probability without specifying firm value and volatility could favor prediction quality. According to recent studies, e.g. Ericsson and Reneby (2004, 2005), the poor prediction quality for structural models is mostly because the traditional approach of Jones et al. (1984) and Ronn and Verma (1986) for estimating these parameters is inadequate. Our results suggest an improvement in performance in the absence of traditional estimation processes. Second, comparing financial models and SVM kernels, Table 4 shows that in four out of six cases the Merton model outperforms the linear SVM, whereas the intensity model is superior to the SVM method in three out of six cases.

Mean Errors and Mean Absolute Errors of prediction are quite low for all three approaches overall, but the Mean Absolute Percentage Error (MAPE) indicates a rather poor out-of-sample fit. Due to the relatively low average premiums (see Table 2) MAPE may partially overstate the prediction errors. Nevertheless, the hypothesis implicitly tested here, that the rating summarizes all relevant credit risk information for a given seniority and region, does not appear to be supported by the data.

5 Remarks and Further Research

The comparison study among structural, reduced-form and SVM methods have yielded some interesting results. The overall prediction quality of all three basic structures is inadequate. It turned out that the Merton model might not be restrictive as it appears, and could outperform a simple version of a reduced-form model. In a few cases, SVM methods were able to achieve competitive results in comparison to financial methods, but they lag behind in most risk-class/kernel combinations. The poor results produced by all three approaches might well be due to the sample design, which employs cross-sectioning with respect to rating, seniority and regions. The most distinctive feature, the ratings, might not be good indicators of the aggregate credit risk. A

Table 3: Out-of-sample Prediction Errors of Merton, Constant Intensity and SVM Methods

Aa Europe, Senior	**ME(bps)**	**MAE(bps)**	**MAPE(%)**	**Estimation**	**Prediction**
Merton	-0.13	4.82	23.82%	7621	3135
Intensity	0.90	5.36	27.64%	7621	3135
SVM Linear	-0.55	5.60	26.77%	7621	3135
SVM Polynomial	-1.05	5.25	24.69%	7621	3135
SVM RBF	-6.87	9.31	41.04%	7621	3135
SVM ERBF	-7.21	9.45	42.57%	7621	3135
Aa N.America, Senior	**ME(bps)**	**MAE(bps)**	**MAPE(%)**	**Estimation**	**Prediction**
Merton	9.13	9.30	43.16%	3475	1571
Intensity	9.87	10.01	46.16%	3475	1571
SVM Linear	3.14	17.75	81.18%	3475	1571
SVM Polynomial	49.85	157.68	657.85%	3475	1571
SVM RBF	-3.33	15.72	68.55%	3475	1571
SVM ERBF	-4.01	14.44	61.89%	3475	1571
A Europe, Senior	**ME(bps)**	**MAE(bps)**	**MAPE(%)**	**Estimation**	**Prediction**
Merton	-2.92	10.19	25.58%	14545	5393
Intensity	0.46	10.20	27.49%	14545	5393
SVM Linear	6.82	12.57	37.17%	14545	5393
SVM Polynomial	6.57	12.51	37.30%	14545	5393
SVM RBF	-0.43	0.81	68.83%	14545	5393
SVM ERBF	-9.51	18.22	41.49%	14545	5393
A N.America, Senior	**ME(bps)**	**MAE(bps)**	**MAPE(%)**	**Estimation**	**Prediction**
Merton	2.00	10.69	25.61%	9046	2093
Intensity	1.98	10.68	25.60%	9046	2093
SVM Linear	-9.70	12.07	23.11%	9046	2093
SVM Polynomial	152.67	152.67	310.13%	9046	2093
SVM RBF	-17.10	18.02	35.06%	9046	2093
SVM ERBF	-4.01	14.44	61.89%	9046	2093
Aa Europe, Subordinate	**ME(bps)**	**MAE(bps)**	**MAPE(%)**	**Estimation**	**Prediction**
Merton	1.38	2.96	11.27%	2094	1168
Intensity	1.56	3.07	11.60%	2094	1168
SVM Linear	-0.06	2.88	11.76%	2094	1168
SVM Polynomial	0.00	2.87	11.70%	2094	1168
SVM RBF	-5.02	7.02	24.90%	2094	1168
SVM ERBF	-5.87	7.66	27.18%	2094	1168
A Europe, Subordinate	**ME(bps)**	**MAE(bps)**	**MAPE(%)**	**Estimation**	**Prediction**
Merton	5.66	8.47	24.49%	3361	1045
Intensity	7.32	9.36	27.00%	3361	1045
SVM Linear	14.82	16.31	42.92%	3361	1045
SVM Polynomial	22.15	23.97	55.33%	3361	1045
SVM RBF	-15.37	21.69	50.36%	3361	1045
SVM ERBF	-17.59	21.90	51.10%	3361	1045

$$\textbf{Mean Error (ME)} = \frac{\sum_{f=1}^{F}\sum_{h=1}^{H} s_f^{theo} - s_{f,h}^{obs}}{F \cdot H}$$

$$\textbf{Mean Absolute Error (MAE)} = \frac{\sum_{f=1}^{F}\sum_{h=1}^{H} \left| s_f^{theo} - s_{f,h}^{obs} \right|}{F \cdot H}$$

$$\textbf{Mean Absolute Percentage Error (MAPE)} = \frac{\sum_{f=1}^{F}\sum_{h=1}^{H} \frac{\left| s_f^{theo} - s_{f,h}^{obs} \right|}{s_{f,h}^{obs}} \times 100}{F \cdot H}$$

s^{theo} is the theoretical CDS premium predicted by the models on day f, where F is the number of available days in the time series. s^{obs} is the observed CDS premium on day f for firm h, where $h = 1 \dots H$ being the number of firms in the prediction sample.

Table 4: Significance Tests for the Difference of Absolute Errors between Merton, Intensity and Linear Kernel SVM in Cross-Sectional Design

	Mean Difference	**t-statistic**	**p-value**	
Aa Europe, Senior				
Merton-Intensity	-0.54	-1.64	0.1011	
Merton-SVM	-0.85	-1.65	0.0998	*
Intensity-SVM	-0.28	-0.6	0.5494	
Aa N.America, Senior				
Merton-Intensity	-0.72	-4.68	< 0.0001	***
Merton-SVM	-8.46	-19.25	< 0.0001	***
Intensity-SVM	-7.75	-24.29	< 0.0001	***
A Europe, Senior				
Merton-Intensity	0.02	0.04	0.9713	
Merton-SVM	-2.32	-2.03	0.0426	**
Intensity-SVM	-2.30	-3.01	0.0026	***
A N.America, Senior				
Merton-Intensity	0.01	0.41	0.6816	
Merton-SVM	-0.49	-0.14	0.8897	
Intensity-SVM	-0.49	-0.14	0.8892	
Aa Europe, Subordinate				
Merton-Intensity	-0.11	-1.88	0.0609	*
Merton-SVM	0.15	0.43	0.6661	
Intensity-SVM	0.26	0.71	0.4804	
A Europe, Subordinate				
Merton-Intensity	-0.79	-2.34	0.0196	**
Merton-SVM	-6.38	-1.69	0.0910	*
Intensity-SVM	-5.91	-1.78	0.0752	*

Mean Difference (bps): Difference of **Absolute Errors** for prediction (Merton-Intensity), (Merton-SVM) and (Intensity-SVM) computed per day per firm.

Absolute Error on day f, for firm $h = \left|s_f^{theo} - s_{f,h}^{obs}\right|$

*** Significance at 99% Level

** Significance at 95% Level

* Significance at 90% Level

similar conclusion is drawn by Frühwirth and Sögner (2006) for the bond market; they claim that any kind of cross-sectioning is inferior to per-firm-based analysis. In this regard, further investigations with the current models should consider time out-of-sample testing as a potential application. This study should look at prediction quality on a firm basis, hypothesizing that the default probability/intensity for a firm is fixed in a given time frame. Altering the models with their recent extensions might be another step in the right direction, as the prediction errors suggest room for further improvement. Moreover, other financial information should be incorporated into the estimation of credit risk models, which can be accomplished by using other instruments (e.g. bonds and stock prices). SVM results could be improved through the application of different kernel/parameter combinations and by inclusion of new variables. It is also highly possible that hybrid approaches consisting of machine learning and financial models could result in superior results. In the end, agreement upon a credit risk modeling framework within the financial modeling practice would clearly facilitate the development of successful market designs.

References

Altman, E. I. and V. M. Kishore (1996): "Almost Everything You Wanted to Know about Recoveries on Defaulted Bonds," *Financial Analysts Journal*, 52, pp. 57–64.

Anderson, R. W. and S. Sundaresan (1996): "Design and Valuation of Debt Contracts," *Review of Financial Studies*, 9, pp. 37–68.

Bakshi, G., D. Madan, and F. X. Zhang (2004): "Investigating the Role of Systematic and Firm-Specific Factors in Default Risk: Lessons from Empirically Evaluating Credit Risk Models," *Forthcoming in Journal of Business.*

Black, F. and J. C. Cox (1976): "Valuing Corporate Securities: Some Effects on Bond Indenture Provisions," *Journal of Finance*, 31, pp. 351–368.

Black, F. and M. S. Scholes (1973): "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, 81, pp. 637–654.

British Bankers' Association (2004): *BBA Credit Derivatives Report 2003/2004*, British Bankers' Association.

Cao, L. and F. E. Tay (2001): "Financial Forecasting Using Support Vector Machines," *Neural Computing and Applications*, 10, pp. 184–192.

Cristianini, N. and J. Shawe-Taylor (2000): *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, UK.

Duffee, G. R. (1999): "Estimating the Price of Default Risk," *Review of Financial Studies*, 12, pp. 197–226.

Duffie, D. and K. J. Singleton (1999): "Modeling Term Structure of Defaultable Bonds," *Review of Financial Studies*, 12, pp. 687–720.

Eom, Y. H., J. Helwege, and J. Huang (2004): "Structural Models of Corporate Bond Pricing: An Empirical Analysis," *Review of Financial Studies*, 17, pp. 499–544.

Ericsson, J. and J. Reneby (2004): "An Empirical Study of Structural Credit Risk Models Using Stock and Bond Prices," *Journal of Fixed Income*, 13, pp. 38–49.

Ericsson, J. and J. Reneby (2005): "Estimating Structural Bond Pricing Models," *Journal of Business*, 78, pp. 707–735.

Ericsson, J., J. Reneby, and H. Wang (2005): "Can Structural Models Price Default Risk?: Evidence from Bond and Credit Derivative Markets," working paper, McGill University.

Frühwirth, M. and L. Sögner (2006): "The Jarrow/Turnbull Default Risk Model: Evidence from the German Market," *European Journal of Finance*, 12, pp. 107–135.

Goldstein, R., N. Ju, and H. Leland (2001): "An EBIT-Based Model of Dynamic Capital Structure," *Journal of Business*, 74, pp. 483–512.

Gündüz, Y., T. Lüdecke, and M. Uhrig-Homburg (2006): "Trading Default Swaps via Interdealer Brokers: Issues towards an Electronic Platform," working paper, University of Karlsruhe.

Houweling, P. and T. Vorst (2005): "Pricing Default Swaps: Empirical Evidence," *Journal of International Money and Finance*, 24, pp. 1200–1225.

International Swaps and Derivatives Association (2004): "ISDA 2004 Operations Benchmarking Survey," working paper, ISDA Inc.

Jarrow, R. A., D. Lando, and S. M. Turnbull (1997): "A Markov Model for the Term Structure of Credit Risk Spreads," *Review of Financial Studies*, 10, pp. 481–523.

Jarrow, R. A. and S. M. Turnbull (1995): "Pricing Derivatives on Financial Securities Subject to Credit Risk," *Journal of Finance*, 50, pp. 53–85.

Jones, E. P., S. P. Mason, and E. Rosenfeld (1984): "Contingent Claims Analysis of Corporate Capital Structures: An Empirical Investigation," *Journal of Finance*, 39, pp. 611–625.

Lando, D. (1998): "On Cox Processes and Credit Risky Securities," *Review of Derivatives Research*, 2, pp. 99–120.

Leland, H. E. (1994): "Corporate Debt Value, Bond Covenants, and Optimal Capital Structure," *Journal of Finance*, 49, pp. 1213–1252.

Leland, H. E. and K. B. Toft (1996): "Optimal Capital Structure, Endogenous Bankruptcy and the Term Structure of Credit Spreads," *Journal of Finance*, 51, pp. 987–1019.

Longstaff, F. A., S. Mithal, and E. Neis (2005): "Corporate Yield Spreads: Default Risk or Liquidity? New Evidence from the Default Swap Market," *Journal of Finance*, 60, pp. 2213–2253.

Longstaff, F. A. and E. S. Schwartz (1995): "A Simple Approach to Valuing Risky Fixed and Floating Rate Debt," *Journal of Finance*, 50, pp. 789–819.

Lyden, S. and D. Saraniti (2000): "An Empirical Examination of the Classical Theory of Corporate Security Valuation," working paper, Barclays Global Investors.

Merton, R. C. (1974): "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates," *Journal of Finance*, 29, pp. 449–470.

Müller, K. R., A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik (1997): "Predicting Time Series with Support Vector Machines," in: *Proceedings of ICANN 1997*, pp. 999–1004.

Ogden, J. P. (1987): "Determinants of the Ratings and Yields on Corporate Bonds: Tests of the Contingent Claims Model," *Journal of Financial Research*, 10, pp. 329–339.

Ronn, E. I. and A. K. Verma (1986): "Pricing Risk-Adjusted Deposit Insurance: An Option-Based Model," *Journal of Finance*, 41, pp. 871–895.

Schwartz, R. A. and R. Francioni (2004): *Equity Markets in Action: The Fundamentals of Liquidity, Market Structure and Trading*, John Wiley & and Sons Inc., Hoboken, New Jersey.

Svensson, L. E. (1994): "Estimating and Interpreting Forward Interest Rates: Sweden 1992-1994," working paper, IMF, 94/114.

Uhrig-Homburg, M. (2002): "Valuation of Defaultable Claims – A Survey," *Schmalenbach Business Review*, 54, pp. 24–57.

Vapnik, V. (1995): *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, NY.

Weinhardt, C., C. Holtmann, and D. Neumann (2003): "Market Engineering," *Wirtschaftsinformatik*, 45, pp. 635–640.

Repeated Decision-making: Incentive Structures and Attitudes to Risk

Jonas Kunze[1], Andreas Geyer-Schulz[1], and Siegfried Berninghaus[2]

[1] Institute of Information Systems and Management, Universität Karlsruhe (TH)
`{jonas.kunze,andreas.geyer-schulz}@em.uni-karlsruhe.de`

[2] Institute of Economic Theory and Operations Research, Universität Karlsruhe (TH),
`berninghaus@wiwi.uni-karlsruhe.de`

Summary. In this paper we give an overview of theoretical and practical results obtained in analyzing the combined factors of repeated decision-making, incentive structures, and attitudes to risk in different fields. Our study is motivated by a recent development in system dynamics in the field of dynamic decision-making. This field investigates the decision-making process of persons in a dynamic context. One characteristic element is the temporal aspect. Decision-making is not static, i.e. a single-shot decision, but done repeatedly. An example of such a dynamic context is the stock management task. In the last years, the stock management task has been investigated in various experiments that applied techniques from the field of experimental economics. In the reported experiments, different incentive schemes were used. This paper focuses on possible effects of the winner-takes-all incentive schemes in repeated decision-making situation. More precisely, we ask what happens to risk attitudes of the subject in such settings. Related results from different disciplines are reported, confirming that the repetition of decisions as well as the winner-takes-all incentive scheme increase the likelihood of risky decisions. Finally, open questions are pointed out.

1 Introduction

How is the decision-making process affected by incentive schemes that calculate a payoff after not only one, but several decisions? This question arose after reviewing studies in dynamic decision-making that treated similar stock management tasks. These tasks include repeated decision-making as the system state changes over time. In our case, the system state includes stock level, number of incoming orders to be served, service-level of the supplier, etc. In order to get an impression about the factors that might influence the decision-making process in such circumstances, we focus on risk attitudes as risk is a well investigated variable in economics. Furthermore, to provide a thorough insight into the effect of incentive schemes on repeated decisions, we restrict our study to the winner-takes-all incentive scheme. So, we reformulate our question to: How are the risk attitudes influenced by the repetition of a decision and by a winner-takes-all incentive scheme?

The paper is structured as follows. In section 2 the motivation for the study is given by describing the different experiments that were done in the field of dynamic decision-making based on the stock management task. In section 3 we focus on the question how the repetition of decisions may effect the risk attitudes of a decision-maker. In section 4 we analyze the relation between the winner-takes-all incentive scheme and risky behavior. Finally, the paper closes with an outlook on further research in the last section.

2 Motivation: System Dynamics and Incentive Structures

The first experimental study of dynamic decision-making in the field of system dynamics was done by Sterman (1989). He used a supply chain simulation ("beer distribution game") with 4 subjects forming a supply chain. Subjects had to decide how many goods (e.g. beer kegs) to order. This decision was repeated in each of the 30 consecutive rounds. After several rounds, the order reached the supplier. If the supplier could fill the order, the goods were shipped and to the subject's stock. Otherwise, a backlog occurred and the order was filled as soon as goods were available.

The overall goal was to minimize the team's transaction costs. For simplicity, only stock costs entered in the calculation. In each round the number of goods in each stock and the number of unfilled orders (backlog) were counted and the actual stock costs were determined. Sterman analyzed the behavior of subjects by estimating parameters for a predefined order heuristic. He argued on the basis of the best-fitting parameters that human decision-makers consistently ignore parts of the supply line, specifically the number of ordered goods that until now have not reached their own stock. With this finding he explained the global phenomenon of increasing upstream order oscillation, which is called the bullwhip effect.

To induce a behavior in which decision-makers try to minimize transaction costs, monetary incentives are used (cf. Smith, 1976). Sterman (1989) used a winner-takes-all incentive scheme, in which the best team, i.e. the team with less transaction costs, gained a prize. Sterman had a high number of accounting errors and little experimental control because he based his analysis on data collected from board game sessions of the beer distribution game. Several years later, Kaminsky and Simchi-Levi (1998) presented a computerized version of the beer game. In their experiment, in which no financial incentives were present, they tested for the effect of shorter lead times and found that the costs produced by the supply chain with shorter lead times were smaller.

Steckel et al. (2004) used an absolute performance based incentive scheme. They wrote: "After the end of the simulation, participant's payments were determined according to the total costs incurred by their entire supply chain. Possible payoffs ranged from \$5 to \$25. The average was \$15." Steckel et al. showed that lower lead times in the order and delivery cycle produced lower costs. They also tested for sharing point-of-sales (POS) data with all decision-makers in the supply chain. Results suggest that depending on the stochastic nature of the external demand function, sharing POS data is helpful in making group decisions.

The studies of Croson and Donohue (2003) and Croson et al. (2004) used a relative performance based incentive scheme. In this type of incentive scheme, the group with the lowest costs gets the highest prize while the group with the highest costs get the lowest prize. The highest and the lowest prize are a fixed amount. The other groups receive a linearly interpolated payment between the highest and the lowest prize based on their relative costs. Croson et al. justify the use of the relative performance based incentive scheme as follows:

> This payment scheme has a number of attractive properties. First, it provides a continuous incentive for participants to earn profit in the game. In contrast to previous experimental implementations in which the highest-earning group won a fixed prize (e.g. Sterman 1989a), each group has an incentive to increase their profits, even if they cannot win first place by doing so. Second, the design discourages collusion among participants to artificially raise the profits of all teams together. Third, the payment represents the benchmarked performance of an integrated supply chain "firm," a performance metric often used in industry.

Croson and Donohue studied the effect of POS data on the bullwhip effect and found that sharing POS data generally reduces the bullwhip effect. Croson et al. also used the bullwhip effect as an efficiency measurement, but focused on the group decision-making process. They identified another behavioral cause of the bullwhip effect that they termed coordination risk. This coordination risk refers to the perceived risk that the other group members will not make optimal decisions. Thus, the player may deviate from optimal decision in order to compensate the anticipated suboptimal decisions of other players.

Summarizing the presented studies, two design aspects are common. First, the dynamic aspect is implemented for experimental settings as discrete time steps. Hence, decision are *repeated*. Second, the payoff is calculated at the end of the session as a function of the transaction costs of all groups. While these aspects remain the same, the exact way of payoff calculation (i.e. the incentive scheme) is different among the studies. In the following sections, we will focus on the payoff calculation method used by Sterman (1989), i.e. the *winner-takes-all incentive scheme*, where only the members of the group with the lowest transaction costs in the session receive a payoff.

3 Repeated Decision-making and Risk

In this section selected topics from the field of economics that deal with repeated decision-making and risk are reviewed. The first part considers the question of whether and how the repetition of some decision-making problems affects the subjects tendency to make risky decisions. The second part examines the interdependence between two subsequent decisions.

When comparing a single risky choice to sequence of risky choices, we start with a well known discussion that began with Samuelson (1963)'s article in which he reported the following situation. Samuelson himself offered a bet to some colleagues at lunch. A coin would be flipped and if the side chosen by his colleague appeared Samuelson would pay him $200. Otherwise his colleague would pay Samuelson $100. His colleague answered that he would decline such a bet, but would accept it if it was repeated 100 times. Samuelson discusses his colleague's behavior and proves that such a behavior cannot be explained by expected utility theory. By induction he shows that if one rejects one favorable bet, at any possible level of wealth, he cannot accept two of such bets, and so on. Finally, he criticizes the term "virtual certainty" that was used by his colleague to justify his choice.

An analysis of the behavior of Samuelson's colleague can be found in Lopes (1981) and Tversky and Bar-Hillel (1983). Lopes argues that Samuelson's colleague's behavior can be explained by choosing gambles that have a greater chance to come out ahead and criticizes the way that expected utility theory predicts the combination of utilities and probabilities. Tversky and Bar-Hillel reject Lopes' criticism by discussing her examples and explaining Samuelson's proof. Furthermore, a psychological analysis of the bet with reference to prospect theory (Kahneman and Tversky, 1979) is presented. Lopes (1996) defends the descriptive power of probability-based decision rules and proposes a dual criterion including decumulative weighting and aspiration level processes. Decumulative weighting is an alternative to expected utility theory and assigns a utility to a possible outcome by weighting the magnitude of the probability as well as the position of the outcome in the set of all available outcomes. Aspiration level processes describe the intent of a decision-maker to favor decisions that have a high probability to return a certain outcome.

Benartzi and Thaler (1995) present an application for Samuelson's colleague's behavior. They investigate the equity premium puzzle that describes the empirical fact that stock portfolios usually outperform bond portfolios. The risk premium in investing is very high; it seems unreasonable to assume a stronger risk aversion for bond investors than for stock investors. Benartzi and Thaler discuss two arguments to explain this, namely *loss aversion* and *mental accounting*. With loss aversion the rejection of one bet while accepting several bets can be explained. Mental accounting (Kahneman, 1984; Thaler, 1985) describes this evaluation process of financial outcomes. Benartzi and Thaler point out that dynamic aggregation rules used in mental accounting introduce a bias if the decision-maker is loss averse. They conclude: "[...] when decision-makers are loss averse, they will be more willing to take risks if they evaluate their performance [...] infrequently". Hence, for a loss averse colleague it is reasonable to reject Samuelson's bet as there is 50% probability of losing, considering it to be an infrequent performance evaluation. However, given 100 repetitions of such a bet the colleague will accept the same odds. Following Benartzi and Thaler, Samuelson's colleague displays *myopic loss aversion.*

In Benartzi and Thaler (1999) four experimental studies are presented in which the myopic loss aversion hypothesis is tested. The subjects in these experiments either had to decide whether to accept a gamble or they had to choose between different gambles. The gambles were presented in different ways. If gambles were not repeated, the gamble with its outcome probabilities was shown. If gambles were repeated a fixed number of times, two presentation formats were used. First, only the gamble that would be repeated with its outcomes probabilities was shown. Second, the distribution of the accumulated outcomes probabilities was explicitly shown. As predicted, subject's myopic loss aversion can be decreased if the distribution of accumulated outcome probabilities is known. This finding is related to what Kahneman and Lovallo (1993) call "narrow framing" that describes a behavior neglecting the aggregation of bets or investments.

Wedell and Böckenholt (1990) present an experimental investigation that supports the hypothesis that repeated gambles reduce the likelihood of preference reversal (Slovic and Lichtenstein, 1983). In preference reversal there are two lotteries. One P-bet that offers a good chance of winning, and a $-bet that offers a lower chance of winning but a higher expected return than the P-bet. If subjects have to price the two lotteries by assigning a minimum selling price, they tend to price the $-bet higher. But, if they have to choose one bet to play, they often prefer the P-bet. This is even true, if these decisions are taken in a within-subject design.

Wedell and Böckenholt compare choice and price decisions of subjects for 1, 10, and 100 repetitions of the underlying P-bet and \$-bet. With increasing repetitions, the incidence of preference reversal decreases. These findings are well explained by the aspiration level of the subjects (Siegel, 1957). If they have to choose, the lotteries are evaluated by comparing the probabilities of reaching their personal aspiration level. Lotteries with a low chance of providing an outcome equal or greater to the aspiration level are filtered out - here: the \$-bet. If the bets are repeated, there is a greater chance for the \$-bet to be selected. Finally, with many repetitions, subjects seem to adopt the long-run perspective in evaluating lotteries.

Next, we focus on the interdependence of successive decisions. We assume that the result of one decision is known to the subject before the next decision has to be made. Hence, the subject gets direct feedback after each decision.

Thaler and Johnson (1990) investigate the effects of prior outcomes on risky choices. They propose quasi-hedonic editing rules in the framework of prospect theory to explain their experimental findings. On the one hand, after gains, subjects are likely to accept risky gambles. This behavioral effect is termed *house money effect*. On the other hand, after losses, two observations are made. First, subjects tend to be risk-averse after an initial loss. Second, if there is any option, no matter how risky, to break-even (i.e. to get back to the initial wealth level) this option is very attractive.

When talking about feedback, we assume that some kind of learning process takes place in the subject. Risky choice experiments mainly present stochastic information about the gambles. Subjects have to make *decisions based on descriptions*. In a comparative study, Weber et al. (2004) show that in all of 226 reviewed choice experiments, stochastic information was given. In contrast, *decisions based on experience*, meaning no stochastic information is provided but through multiple trials the subject is able to perceive the character of the probability distribution, is poorly investigated. A study by Hertwig et al. (2004) focuses on the different decision modi. They propose that *decisions based on experience* underweight rare events because of limited information search and recency effects.

Ansic and Keasey (1994) present the results of a pilot study in which they tested for significant correlation between subjects' current and past decisions. Furthermore, they investigated whether the decisions have a stationary character or show a trend over time. The experiment consisted of a financial market simulation in which subjects had to make approximately 35 buy/sell decisions for one share of stock based on a random series of price. These decisions were coded into a time series of observations z describing the change in terms of percentage - positive or negative - from the subject's cash stock to his share stock. Most subjects showed autoregression in their time series for lags of up to 10 periods during which the influence of the last decision continued to affect decision-making. Forty percent of the subjects showed a non-stationary behavior. From these results it can be concluded that past decisions have a great influence on current decisions which can be clearly observed in often repeated decisions. Open issues of the study include learning aspects and the influence of incentive systems that will be discussed in the next section.

There is experimental evidence that people adopt a long-run perspective if a gamble is offered repeatedly (Wedell and Böckenholt, 1990; Benartzi and Thaler, 1999; Samuelson, 1963). Especially, if the resulting payoff distribution is explicitly shown to the subjects (Benartzi and Thaler, 1999; Lopes, 1996). These findings suggest that risk neutral decision behavior is promoted by continued repetition of a game. But if we change to a piecewise repetition (e.g. instead of a 1000 times repetition, ten times 100 repetitions) subject's behavior changes (cf. portfolio evaluation discussed in Benartzi and Thaler 1995). Furthermore, if we let people decide for each of the set of repetitions separately, the decisions tend to vary (Thaler and Johnson, 1990) and show a high interdependence (Ansic and Keasey, 1994). These findings are based on the fact that outcome feedback is provided after each decision. An open question is, how behavior might be influenced if the outcome feedback were misleading (e.g. not reflecting precisely the actual situation). The misleading effect of delaying feedback is of special interest.

Resuming this section, repetition may lead to risk neutral behavior if the subject is able to integrate the decisions. Otherwise, it is likely that prior gains (house money effect) or losses with the chance to break even (aspiration level) induce risk-taking behavior.

4 Risk and the Winner-takes-all Incentive Scheme

Let us now focus on the incentive scheme. The winner-takes-all incentive scheme is an extreme form of a rank-based incentive scheme. In the winner-takes-all version, only the best player (winner) receives a prize. To determine the best player, each player is assigned some *performance measure*. Next, this performance measure is maximized or minimized to determine the best player. In the stock management task the performance measure is the level of transaction costs and the best player is the one minimizing them. We now consider fields, race models and group incentive, where winner-takes-all incentive schemes are investigated.

A race is a competition in which a prize is awarded to the winner. The performance measure is the time and the winner is the one minimizing it. In economic contexts the patent race embodies an application for models describing races. In these models, players choose strategies that are associated with costs. Then, after some time an invention is made and a patent is granted, unless another player already patented the invention. Patent race models treat different aspects of the research process. First, there is uncertainty in the time that is needed until an invention occurs. Loury (1979), Lee and Wilde (1980) treat this uncertainty aspect. Second, there might be strategic interaction between players. Here, inventing process is split up into several parts. Knowing the amount of progress made in the past by each party it is possible to speak of a leader and a follower. In Fudenberg et al. (1983), Harris and Vickers (1985a), and Harris and Vickers (1985b) race models that focus on the strategic aspect are described. It is shown that leaders in a race have a nearly competition-free situation and followers do not expend great effort as any effort will be outperformed by the leader. These finding are confirmed by Harris and Vickers (1987) by using models that include uncertainty as well as strategic interaction. Third, riskiness in terms of mean preserving spread is associated with the different strategies. Dasgupta and Maskin (1987) compare social optimal strategies of player with strategies under competition. The main finding regarding the risk aspect is that the strategies of players under competition are too risky in comparison to the social optimum. This is explained by the character of the incentive scheme that while society is interested in at least one advance, competitors focus on being the first one.

Winner-takes-all incentive scheme are also often used in the field of team production. Dickinson and Isaac (1998) compare between absolute and relative rewards in team production assuming asymmetric endowment of players. The highest contributing player, in absolute or relative terms, is given a price. Experimental results show that the existence of a prize increases the contribution level dramatically, but in case of the absolute reward calculation after some rounds only the best endowed player contributes a high level and gains the prize.

Overall, the winner-takes-all incentive scheme pushes players to perform well. Depending on the model, or high efforts are made (Loury, 1979; Lee and Wilde, 1980; Dickinson and Isaac, 1998) or risky strategies are chosen (Dasgupta and Maskin, 1987) to achieve this goal. If asymmetries are present, efforts are reduced if the player is not the best performing one (Fudenberg et al., 1983; Harris and Vickers, 1985a,b).

5 Outlook

If decisions are integrated, decision-makers tend to adopt a long run perspective, hence, risk neutrality is induced. But if decisions are made separately, other effects occur. Focusing on each decision may lead to higher risk averseness (myopic loss aversion). Contrarily to this, perfect outcome feedback often induces risk taking behavior that could be explained by the house money effect and by the aspiration level theory. If the incentive scheme has a winner-takes-all structure, risky decisions are likely. Overall, there are opposed effects referring to the risk attitudes. In order to make more precise statements, several open issues exist. First, a quantitative analysis of the discussed settings is needed. Second, the type of outcome feedback seems to be of great importance, especially non-perfect feedback, e.g. delayed or partial. Third, other incentive schemes should be analyzed. Fourth, having come to a clear result about risk attitudes in repeated decisions with a winner-takes-all incentive scheme, the consequences for the design of dynamic decision-making experiments need to be developed.

References

Ansic, D. and K. Keasey (1994): “Repeated decisions and attitudes to risk,” *Economics Letters*, 45, pp. 185–189.

Benartzi, S. and R. H. Thaler (1995): “Myopic Loss Aversion and the Equity Premium Puzzle,” *The Quartely Journal of Economics*, 110(1), pp. 73–92.

Benartzi, S. and R. H. Thaler (1999): “Risk Aversion or Myopia? Choices in Repeated Gambles and Retirement Investments,” *Management Science*, 45(3), pp. 364–381.

Croson, R. and K. Donohue (2003): “Impact of point of sale (POS) data sharing on supply-chain management: An experimental approach,” *Productions and Operations Management*, 12(1), pp. 1–11.

Croson, R., K. Donohue, E. Katok, and J. Sterman (2004): “Order Stability in Supply Chains: Coordination Risk and the Role of Coordination Stock,” working paper, Massachusetts Institute of Technology, Engineering Systems Division.

Dasgupta, P. and E. Maskin (1987): “The Simple Economics of Research Portfolio,” *The Economic Journal*, 97(387), pp. 581–595.

Dickinson, D. L. and R. M. Isaac (1998): "Absolute and Relative Rewards for Individuals in Team Production," *Managerial and Decision Economics*, 19(4/5), pp. 299–310.

Fudenberg, D., R. Gilbert, J. Stiglitz, and J. Tirole (1983): "Pre-emption, Leapfrogging and Competition in Patent Races," *European Economic Review*, 22, pp. 3–31.

Harris, C. and J. Vickers (1985a): "Patent Races and the Persistence of Monopoly," *The Journal of Industrial Economics*, 33(4), pp. 461–481.

Harris, C. and J. Vickers (1985b): "Perfect Equilibrium in a Model of a Race," *The Review of Economic Studies*, 52(2), pp. 193–209.

Harris, C. and J. Vickers (1987): "Racing with Uncertainty," *The Review of Economic Studies*, 54(1), pp. 1–21.

Hertwig, R., G. Barron, E. U. Weber, and I. Erev (2004): "Decisions From Experience and the Effect of Rare Events in Risky Choice," *Psychological Science*, 15(8), pp. 534–539.

Kahneman, D. (1984): "Choices, Values and Frames," *American Psychologist*, 39(4), pp. 341–350.

Kahneman, D. and D. Lovallo (1993): "Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking," *Management Science*, 39(1), pp. 17–31.

Kahneman, D. and A. Tversky (1979): "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47(2), pp. 263–292.

Kaminsky, P. and D. Simchi-Levi (1998): "A new computerized beer game: A tool for teaching the value of integrated supply chain management," in: H. Lee and S. M. Ng (eds.), *Global Supply Chain and Technology Management*, *POMS Series in Technology and Operations Management*, vol. 1, pp. 216–225.

Lee, T. and L. L. Wilde (1980): "Market Structure and Innovation: A Reformulation," *The Quarterly Journal of Economics*, 94(2), pp. 429–436.

Lopes, L. L. (1981): "Decision making in the short run," *Journal of experimental psychology: Human Perception and Performance*, 7(5), pp. 377–385.

Lopes, L. L. (1996): "When Time Is of the Essence: Averaging, Aspiration, and the Short Run," *Organizational Behavior and Human Decision Processes*, 65(3), pp. 179–189.

Loury, G. C. (1979): "Market Structure and Innovation," *The Quarterly Journal of Economics*, 93(3), pp. 395–410.

Samuelson, P. A. (1963): "Risk and Uncertainty: A Fallancy of Large Numbers," *Scientia*, 98, pp. 108–113.

Siegel, S. (1957): "Level of Aspiration and Decision Making," *Psychological Review*, 64, pp. 253–262.

Slovic, P. and S. Lichtenstein (1983): "Preference Reversals: A Broader Perspective," *The American Economic Review*, 73(4), pp. 596–605.

Smith, V. L. (1976): "Experimental Economics: Induced Value Theory," *American Economic Review*, 66(2), pp. 274–279.

Steckel, J. H., S. Gupta, and A. Banerji (2004): "Supply Chain Decision Making: Will Shorter Cycle Times and Shared Point-of-Sale Information Necessarily Help?" *Management Science*, 50(4), pp. 458–464.

Sterman, J. D. (1989): "Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Experiment," *Management Science*, 35(3), pp. 321–339.

Thaler, R. H. (1985): "Mental Accounting and Consumer Choice," *Marketing Science*, 4(3), pp. 199–214.

Thaler, R. H. and E. J. Johnson (1990): "Gambling with the House Money and Trying to Break Even: The Effects of Prior Outcomes on Risky Choice," *Management Science*, 36(6), pp. 643–660.

Tversky, A. and M. Bar-Hillel (1983): "Risk: The Long and the Short," *Journal of experimental psychology: Human learning, Memory, and Cognition*, 9, pp. 713–717.

Weber, E. U., S. Shafir, and A.-R. Blais (2004): "Predicting Risk Sensitivity in Humans and Lower Animals: Risk as Variance or Coefficient of Variation," *Psychological Review*, 111(2), pp. 430–445.

Wedell, D. H. and U. Böckenholt (1990): "Moderation of Preference Reversal in the Long Run," *Journal of Experimental Psychology: Human Perception and Performance*, 16(2), pp. 429–438.

Predictive Power of Markets: A Comparison of Two Sports Forecasting Exchanges

Stefan Luckner[1], Christof Weinhardt[1], and Rudi Studer[2]

[1] Institute of Information Systems and Management, Universität Karlsruhe (TH)
{luckner,weinhardt}@iw.uni-karlsruhe.de

[2] Institute of Applied Informatics and Formal Description Methods (AIFB), Universität Karlsruhe (TH)
studer@aifb.uni-karlsruhe.de

Summary. Forecasting markets are a promising approach for predicting future events. Prior experience demonstrates that real-money as well as play-money markets predicted future events at a remarkable accuracy. Experimental economists most probably would insist that monetary risk is required in order to obtain valid conclusions about economic behavior. On the other hand, there is evidence that monetary incentives do not necessarily increase performance. In this paper, we study the impact of monetary investments by comparing predictions from the play-money market STOCCER to those of the real-money market BLUEVEX. Perhaps surprisingly, the play-money market STOCCER performs better than the real-money market BLUEVEX. However, we believe it is speculative to assume that play money was the only reason for the good prediction derived from market prices in STOCCER. We discuss several differences between the two sports forecasting exchanges as well as their impact on the market outcome and thus motivate more elaborate future experiments.

1 Introduction

Forecasting markets are a promising approach for predicting future events. The basic idea of a forecasting market is to trade virtual stocks whose final value is tied to the outcome of a particular future event. Once the outcome is known, each stock receives a payoff. Market prices can thus be interpreted as predictions of the probability of those future events. In information efficient markets, prices represent all available information about the participants valuations at any time (Fama, 1970). The results of recent studies on forecasting markets are encouraging. The Iowa Electronic Market (IEM) for predicting the outcome of the presidential elections in 1988 was the first political stock market (Forsythe et al., 1992). By then, the accuracy of the prediction was amazing and much better than traditional polls. Since that time, political stock markets have been widely used as an alternative to polls and initially seemed to be the miracle cure in psephology. Apart from political stock markets, the idea behind forecasting markets based on the efficient market hypothesis and Hayek's theories about the information efficiency of markets (Hayek, 1945) has also been used in various other fields of application like in market research or business forecasting in general (Spann and Skiera, 2003, 2004). Lately, forecasting markets are also used in order to predict the outcome of sports events.

Recent forecasting markets were mostly based on a similar design, although it is known that the market design heavily impacts the market outcome. Even minor changes in the design can totally sway the behavior of market participants. Thus, various design issues in such markets deserve further study. Our work focuses on designing markets along the lines of market engineering. Market engineering is the deliberate design of market institutions in their entirety: it involves embedding the microstructure, the infrastructure as well as the business and governance structure of a market (Weinhardt et al., 2003). The focal point of this work is one particular aspect of forecasting markets, namely the effect of real money instead of play money investments by market participants on the predictive power.

Experimental economists most probably would insist that monetary risk is required in order to obtain valid conclusions about economic behavior. Payments based on the participants' performance are usually intended to provide incentives for rational - or at least well considered - decision making. On the other hand, there is evidence that monetary incentives do not necessarily increase performance (Gneezy and Rustichini, 2000). Prior experience in the field of forecasting markets demonstrates that real-money as well as play-money markets predicted future events at a remarkable accuracy. In this paper, we study the impact of monetary investments by comparing the predictions from the play-money market STOCCER[1] to those of the real-money market BLUEVEX[2].

The following section describes our real-world online experiment. We then discuss first results concerning the predictive accuracy of real-money vs. play-money forecasting markets in section 3. In our study, STOCCER performed better than BLUEVEX. Therefore we also speculate why using play-money markets could potentially result in better predictions than real-money markets. In section 4, we discuss several differences between the two sports forecasting exchanges as well as their impact on the market outcome and conclude that it is still speculative to assume that play money was the prime reason for the good prediction of STOCCER. Thus, we motivate more elaborate future experiments before concluding with some implications of our results on practice and giving an outlook on future work.

2 Experimental Setup

We compare two sports forecasting exchanges that both operated a market for the German Soccer League in 2005: STOCCER and BLUEVEX. Both tried to predict the top team after the first half of the 05/06 season and were open to the public 24 hours a day, 7 days a week. BLUEVEX is run by a German operator of online sports contests and requires real money investments from market participants. Trading profits are hence paid out in monetary form. Moreover, BLUEVEX charges a small fee on each transaction. Based on our own market platform we set up another forecasting market for the German Soccer League which we called STOCCER. In contrast to BLUEVEX, we decided to operate STOCCER as a play-money market. Instead of investing real money every market participant had an initial endowment of 100.000 virtual currency units. The only extrinsic incentive for traders to join the market and reveal their expectations was a ranking of their user names.

[1] http://www.stoccer.com

[2] http://www.bluevex.de

Both forecasting exchanges offered similar virtual stocks. In case of BLUEVEX, a soccer team's virtual stock was valued at 10 Euro if the team was the top team after the first half of the 05/06 season, and 0 Euro otherwise. In case of STOCCER, the top team's virtual stock was valued at 100 virtual currency units; all other stocks were valued at 0. STOCCER and BLUEVEX also had many commonalities in their financial market design. Both offered a continuous double auction in combination with limit orders. However, the primary markets differed. While stocks were issued by the market operator in case of BLUEVEX, we decided to offer so called portfolios in our forecasting market STOCCER. A portfolio contains one share of every possible outcome, i.e. every stock that is traded in the market. The portfolio price was 100 virtual currency units and thus corresponded to the redemption for correctly predicting the outcome. Buying and selling portfolios was possible at all times. Table 1 summarizes the comparison of STOCCER and BLUEVEX.

Table 1: Design of STOCCER and BLUEVEX

Design Issue	**BLUEVEX**	**STOCCER**
Forecasting goal	• Top team of the German Soccer League after first half of the 05/06 season • Payment of 10 Euro for every stock of top team • Open to public	• Top team of the German Soccer League after first half of the 05/06 season • Payment of 100 virtual currency units for every stock of top team • Open to public
Incentives for information revelation	• Players must deposit money • Performance-based monetary reward	• Endowment 100.000 virtual currency units per participant • Ranking as nonmonetary reward
Financial market design	• Primary market: emitted by BLUEVEX • Trading mechanism: continuous double auction • Trading times: 24 hours a day, 7 days a week • Order type: limit orders • Trading fee: 0.05 Euro per traded stock	• Primary market: portfolio trading • Trading mechanism: continuous double auction • Trading times: 24 hours a day, 7 days a week • Order type: limit orders • Trading fee: none

Our real-world experiment started on 31 October 2005 and ran until the end of the first half of the season on 19 December 2005. Stock prices were recorded every day at 1 p.m. during this period of 50 days. Traders on both forecasting exchanges did not know that the market prices resulting from their transactions were used for this experiment. On average, we had 50 transactions per day from a total number of 135 registered users on STOCCER. The daily number of transactions for BLUEVEX was not available, but based on the total trading volume over the period of 50 days we assume that it was smaller than in STOCCER. Considering the large number of registered users on BLUEVEX this is rather surprising.

3 Results

As expected, both markets exhibited significant predictive power. Starting from the first day of our experiment, virtual stocks of the soccer team "FC Bayern München" were traded at higher prices than any other stock. In both forecasting markets, the probability of "FC Bayern München" being the top team was never valued lower than 70%. In the end, the "FC Bayern München" was the top team after the first half of the 05/06 season.

To determine the forecast accuracy, we computed the average absolute deviation (AAD) of the market prices of all stocks at a given time from the actual payoff for the corresponding stocks after the market closed. Measured by the AAD, a comparison of the forecast accuracy of BLUEVEX and STOCCER is depicted in figure 1.

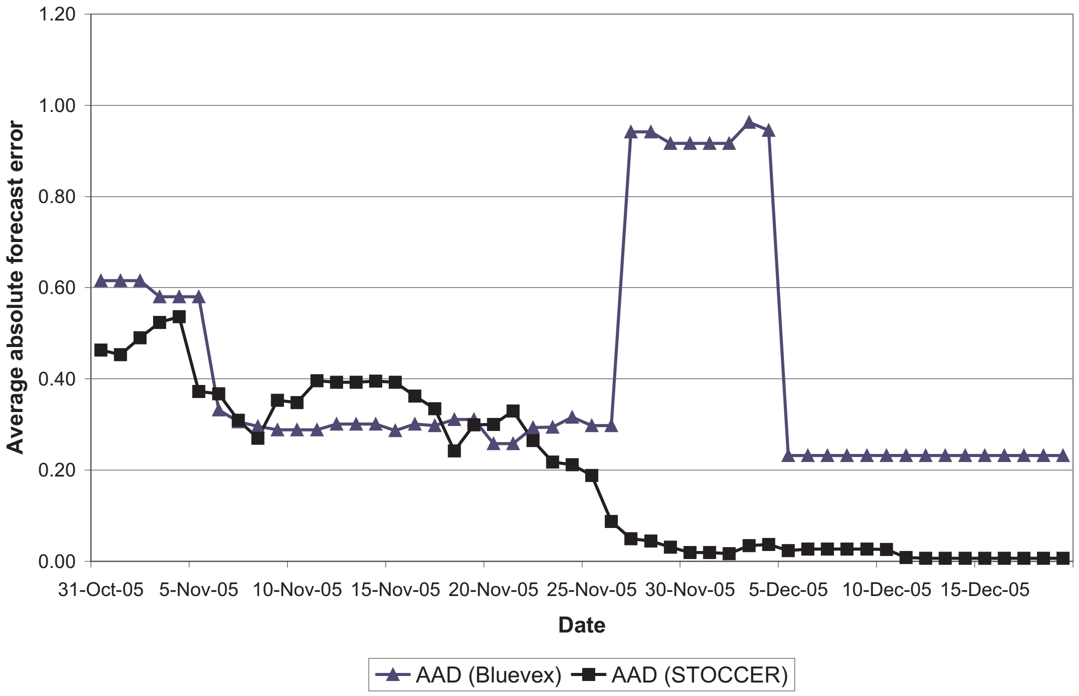


Figure 1: Forecast accuracy of STOCCER and BLUEVEX through time

In 37 out of 50 days, the AAD was lower in case of STOCCER. The mean value of the AAD reached at 0.42 in BLUEVEX and at 0.20 in STOCCER. We thus argue that in case of the German Soccer League our play-money market STOCCER performed better than the real-money market BLUEVEX. At least in case of STOCCER, the graph also shows how the accuracy of the forecast improves, i.e. the AAD declines, as information is revealed and absorbed by the market in the course of time. Surprisingly, the AAD remains at a rather high level in case of BLUEVEX. We will discuss possible reasons for this in section 4.

From November 27 to December 4, the AAD observed in BLUEVEX was extremely high. This results from a single transaction. One trader bought the stock of the team "Borussia Dortmund" at a price of 9.99 Euro despite the fact that this team did not have a reasonable chance to become the top team. Afterwards, "Borussia Dortmund" was not traded for a couple of days and thus stayed at a price level of 9.99 Euro. When neglecting this single transaction, the AAD of BLUEVEX can be reduced to a mean value of 0.30. Nevertheless, STOCCER is still performing better in 37 out of 50 days. The development of the AAD over time when omitting the afore-mentioned transaction is shown in figure 2.

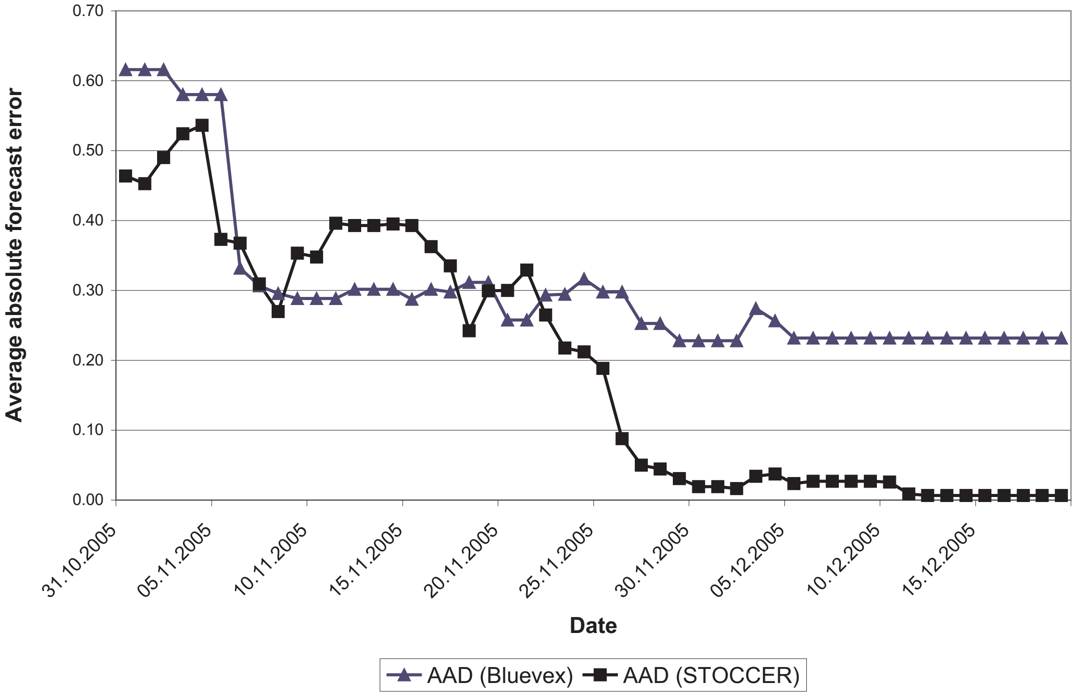


Figure 2: Forecast accuracy of STOCCER and BLUEVEX through time (omitting one single transaction of the stock "Borussia Dortmund" in BLUEVEX)

This raises the question, why the prediction derived from the market prices of the play-money market was better compared to the prediction based on the market prices of the real-money market. There are arguments to suggest that markets where traders risk their own money perform better than play-money markets. First, using real money can be seen as an incentive for participants to seek accurate information and trade reasonably. After all, traders risk their own money in such a market. Second, real-money markets provide an incentive for truthful revelation. It's in the traders interest to reveal their expectations and trade accordingly. Third, traders will be willing to bet more on forecasts they are more confident about. This means that well-informed traders tend to invest more money in such a market. Theory thus suggests that real money may better motivate information discovery (Servan-Schreiber et al., 2004).

On the other hand, there are also reasonable arguments in favor of play-money markets. In contrast to real-money markets, a player's market power does not depend on his financial power in real life, but rather on his success in the history of a forecasting market. Successful traders have a higher deposit value at their disposal and can for that reason exert more influence on the market than unsuccessful traders. Moreover, material incentives such as prizes or immaterial incentives like rankings or the announcement of the most successful traders on web sites or in newsletters can also provide incentives for truthful revelation of traders expectations. Such incentives oftentimes suffice to motivate intense trading. Finally, the intrinsic motivation of the traders should not be underestimated. Play instinct, ambition to be first in a ranking or just fun may oftentimes be enough to attract traders revealing their expectations.

Recapitulating, in order to achieve the highest possible accuracy knowledgeable traders have to participate. The key question thus is how to motivate those traders. Furthermore, traders need an incentive to reveal their information. While traders are by default trying to do their best in real-money markets they can also be motivated by rankings, prizes or the like in play-money markets

4 Discussion

In our case the play-money market STOCCER performed better than the real-money market BLUEVEX. In an earlier study, Servan-Schreiber et al. (2004) found that there was no statistically significant difference between the real-money market TradeSports and the play-money market NewsFutures. Rosenbloom and Notz (2006) however, found TradeSports to be significantly more accurate than NewsFutures for non-sports events. In case of NFL games, they produced conclusions consistent with those from Servan-Schreiber et al. (2004). However, both studies do not consider any other differences apart from the use of real money or play money in their comparison of forecasting markets. Although the two forecasting markets we compared are quite similar, they are not identical in their market design. In our opinion, a key difference is indeed that one uses real money while the other uses play money. But was the use of play money really the reason for the success of STOCCER? We claim this remains an open question because BLUEVEX and STOCCER also differ in some other aspects:

Number of users

In case of STOCCER, we had a total number of 135 registered users and 50 transactions on average per day. However, we do not have any information on the number of users or the number of transactions in BLUEVEX. Based on the total trading volume over the period of 50 days we assume that both were smaller than in STOCCER. Therefore, the question arises whether the higher trading volume in STOCCER was the reason for greater predictive accuracy. Data collected by Forsythe et al. (1999) indicates that a small number of traders only is necessary to drive forecasting markets to efficient outcomes. According to them, the performance of prediction markets heavily depends on a rather small number of well-informed and motivated traders, so called marginal traders. Since the real-money market BLUEVEX is used by many experienced traders (e.g. markets are almost at all times arbitrage free), we suppose that the smaller number of users in BLUEVEX cannot explain the inferior predictive accuracy. On the other hand, we observe that trading in BLUEVEX comes to an end quite some time before the outcome of the event is known. As a consequence, new information is not reflected in the market prices and the forecast error remains at a higher level compared to STOCCER. A larger number of users with more diverse expectations could possibly have increased the trading activity during the last match days.

Transaction fee

BLUEVEX charges a small fee on each transaction. Costs of trading stocks of course also impact the market price. On the one hand, the impact of transaction fees in BLUEVEX can be neglected because implicit transaction costs in terms of spread from our experience by far exceeded explicit transaction costs in terms of fees. On the other hand, transaction fees can be another reason for reduced trading activity. Especially during the last match days many teams had a very low probability of becoming the top team. In STOCCER, these stocks were traded at prices close to zero. Transactions fees in BLUEVEX restrain users from trading such penny stocks because a fee of 0.05 Euro per stock is relatively high in relation to the stock price. As a result, the last price of several unsuccessful teams remains at a higher level compared to STOCCER and thus increases the forecast error.

Portfolio trading

As explicated in section 2, the primary markets differed in the sense that STOCCER offered so called portfolios. Since the portfolio price corresponds to the redemption for correctly predicting the outcome, risk-free portfolio trading was possible in STOCCER. As a consequence, the sum of the market prices of all stocks was always close to the redemption for the stocks of the winning team. We observed that in case of BLUEVEX the sum of the market prices was oftentimes much higher than the redemption. This is rather surprising because BLUEVEX allows for short selling. We tried to reduce this effect by scaling the market prices to the redemption. The mean value of the AAD then reached at 0.47 in BLUEVEX and at 0.22 in STOCCER. One conclusion could be that the better performance of STOCCER cannot originate from the portfolio trading functionality. However, the availability of portfolios in STOCCER has a second effect. It offers the possibility of arbitrage trading and consequently increases market liquidity - and the illiquidity seems to be one major shortcoming in case of BLUEVEX.

In the end, all of the above-mentioned differences have to be taken into consideration when comparing STOCCER to BLUEVEX as well as any other two forecasting markets. It is thus speculative to assume that play money was the reason for the better performance of STOCCER compared to BLUEVEX. Both previous studies, the one by Servan-Schreiber et al. (2004) and the other by Rosenbloom and Notz (2006), have the same weak point. What remains to be done is an experiment where using play money instead of real money is the only difference between two markets. This is part of our future work. The other differences like portfolio trading and transaction fees need to be studied in separate experiments.

5 Conclusion and Outlook

In this paper we described a real-world experiment for comparing the forecast accuracy of the real-money market BLUEVEX and the play-money market STOCCER. We observe that our play-money market performed better than the real-money market and speculate about possible reasons for the somewhat surprising result. By comparing the design of both forecasting markets we find that there are other differences in the market design apart from using play money instead of real money. In our opinion, it is thus speculative to assume that play money was the reason for the better performance of STOCCER compared to BLUEVEX. Earlier studies have the same weak point.

Up to now, there have only been two studies that examined whether real-money markets perform better than play-money markets. Servan-Schreiber et al. (2004) found that there was no statistically significant difference between TradeSports and NewsFutures whereas Rosenbloom and Notz (2006) found the real-money market TradeSports to be significantly more accurate than the play-money market NewsFutures for non-sports events. Considering both studies, we believe that the impact of real money vs. play money remains an open question in the field of forecasting markets. Our real-world experiment gives evidence that real-money markets do not necessarily perform better than play-money markets. In a next step, we will set up a field experiment for getting a larger number of observations. We want to study data collected from a larger number of comparable real-money as well as play-money markets for single matches of the FIFA World Cup 2006.

Comparing the forecast accuracy of real-money to play-money markets also has practicable implications. Due to technical and also legal reasons it is much easier to set up a play-money market. In many countries, operating a real-money market is outlawed. From a more technical perspective, money inflows as well as outflows have to be handled and system failures may not occur. We thus conclude that the comparison of play-money to real-money markets deserves further study.

References

Fama, E. (1970): "Efficient capital markets: A review of theory and empirical work," *Journal of Finance*, 25, pp. 383–417.

Forsythe, R., F. Nelson, and J. Neumann, G. Wright (1992): "Anatomy of an Experimental Political Stock Market," *American Economic Review*, 82, pp. 1142–1161.

Forsythe, R., T. A. Rietz, and T. Ross (1999): "Wishes, expectations and actions: a survey on price formation in election stock markets," *Journal of Economic Behavior and Organization*, 39, pp. 83–110.

Gneezy, U. and A. Rustichini (2000): "Pay Enough Or Don'T Pay At All," *The Quarterly Journal of Economics*, 115, pp. 791–810.

Hayek, F. (1945): "The Use of Knowledge in Society," *American Economic Review*, 35, pp. 519–530.

Rosenbloom, E. S. and W. W. Notz (2006): "Statistical Tests of Real-Money versus Play-Money Prediction Markets," *Electronic Markets - The International Journal*, 16.

Servan-Schreiber, E., J. Wolfers, D. Pennock, and B. Galebach (2004): "Prediction Markets: Does Money Matter?" *Electronic Markets - The International Journal*, 14.

Spann, M. and B. Skiera (2003): "Internet-Based Virtual Stock Markets for Business Forecasting," *Management Science*, 49, pp. 1310–1326.

Spann, M. and B. Skiera (2004): “Einsatzmöglichkeiten virtueller Börsen in der Marktforschung,” *Zeitschrift für Betriebswirtschaft (ZfB)*, 74 (EH2), pp. 25–48.

Weinhardt, C., C. Holtmann, and D. Neumann (2003): “Market Engineering,” *Wirtschaftsinformatik*, 45, pp. 635–640.

Personality in Social Networks. A Theoretical Overview

Cora Schaefer[1], Andreas Geyer-Schulz[1], and Siegfried Berninghaus[2]

[1] Institute for Information Systems and Management,
Universität Karlsruhe (TH)
{cora.schaefer,andreas.geyer-schulz}@em.uni-karlsruhe.de

[2] Institute for Economic Theory and Operations Research,
Universität Karlsruhe (TH)
siegfried.berninghaus@wiwi.uni-karlsruhe.de

Summary. In this paper the role of personality in social network research is reviewed and analyzed. To begin with, basic concepts and the operating paradigm of social network analysis are introduced. Next, theories and research within other paradigms concerning the role of the individual in forming his social environment are presented along with further empirical findings. Different limitations of the empirical evidence available to date are discussed. This review shows that personality aspects play an important role in social networks; however, today their impact to the structure of our social worlds constitutes a blank spot on the map of social network research. The influence of personality in the formation of social networks should therefore be systematically investigated.

1 Introduction

Some people seem to know the whole world while others stick to their small circle of friends. It is self-evident that the persons with wide-ranging networks spread information quickly, but what do others make of the information and how is it evaluated? The influence of a particular bit of information is determined by the status and credibility of the messenger. So it is not who or how many you know, but rather how they know *you*. The size and scope of one's network as well as the individual position within the network varies from person to person. The identification of persons occupying central positions (and therefore being able to influence others) is of critical importance in many domains, including market analysis, marketing and the dynamics of organizational change processes.

For a long time, social network research limited itself to the description and investigation of the consequences of communication structures in social networks. Antecedents, particularly individual variables like extraversion or social skills, were not considered. It is precisely these psychological factors though, that influence the behavior and therefore the role of a person in the development of a network structure, and ,in turn, the development of the network as a whole.

Uses of Social Network Insights.

Currently, social network analysis is applied to all kinds of social networks in the hope of gaining a deeper or readily accessible insight into their functioning. A wealth of research treats organizational networks on the individual, unit-based and organizational level (for recent reviews see Brass et al. (2004) and Borgatti and Foster (2003)). The spread of information through customer networks, for example, has become an important topic in market engineering and customer relationship management. Managing this information flow allows firms to address only a preselected number of persons in their marketing campaigns, yet reach many (e.g. Domingos and Richardson, 2001). For this to be accomplished the "right" persons to target, i.e. the influential and central individuals within preferably large networks, have to be found. Other areas, such as computer-supported networks (Chang et al., 2002; Haythornthwaite, 2002) or market analysis, now come into focus.

2 Social Network Research

Basic principles.

Social network analysis investigates the relationship patterns among any set of humans. A social network describes the pattern of relationships in which actors interact with one another and through the emergent structures they are in turn influenced in their behavioral options. Formally, a social network is defined as a "set of actors and the ties among them" (Wassermann and Faust, 1994, p. 9). The tie may refer to any kind of relationship, e.g. economic, transactional, politic or informational, and is "limited only by a researcher's imagination" in terms of content (Brass et al., 2004). The unit or actor is termed node. When focusing on one single actor and his personal network, that actor is called "ego" and his contacts "alters". The sum of all ties among the nodes constitutes an "ego-network" (Wassermann and Faust, 1994).

The methods of social network analysis allow one to either study the network structure as a whole or analyze individual positions within the network or different parts of it. Several measures are used to describe a personal network: first, the number of ties ego has constitutes the size of his personal network. Another interesting variable is the scope or range of his network, which denotes how many persons he can reach directly or indirectly through his alters. Within a network, status, a sociological concept, describes the position a person occupies in the structure of social relations relative to others. Status is by definition a relative concept. A status measure is therefore only valid in comparative interpretation (Hartfiel and Hillmann, 1982). Network analysis indicates the status of an actor by means of a centrality measure, an index derived from his position in relation to the position of the other actors in the network. A simple status measure can be deduced from the direct in- and outgoing connections of one actor to others in the network (degree centrality). This index can be further differentiated by distinguishing the ingoing and outgoing connections (indegree and outdegree centrality, respectively). The more direct connections an actor has, the higher his status. The most central actor, then, is the one with the most direct connections. More sophisticated status measures not only consider ego's direct connections, but also the indirect connections within the network.

Paradigm.

One of the main goals of social network research is the investigation of the reasons for and influencing factors of network evolution and decay (Schweizer, 1989). Yet the wealth of the literature addresses the outcomes of network variables while network antecedents are rarely studied. Borgatti and Foster see one reason for this in the social network research being a relatively young research area which had yet to establish itself by showing the importance of network outcomes (Borgatti and Foster, 2003).

In particular, individual level antecedents of network formation were seldom treated (Borgatti and Foster, 2003; Mehra et al., 2001; Klein et al., 2004). Studies examining the causalities of network characteristics often fail to follow the structuralist heritage explaining network evolution and change in terms of personalities and other latent variables of the actors (Borgatti and Foster, 2003). The structuralist approach of social network analysis is concerned solely with the relations between actors and the network influence on the actors. Thus, within the structuralist paradigm, individual behavior is assumed to be first and foremost a product of the network environment and embeddedness of the actor (Jansen, 2003). Individual differences are seen as a consequence of networks and can therefore be dismissed: “instead of analyzing individual behaviors, attitudes, and beliefs, social network analysis focuses its attention on social entities or actors in interaction with one another ...” (Galaskiewicz and Wassermann, 1994). Wassermann and Faust (1994, p. 5) clarify the relational approach further: “the unit of analysis in network analysis is not the individual, but an entity consisting of a collection of individuals and the linkages among them.” Thus, it was long a central supposition that social network scholars need not or rather should not consider individuals nor variables on the individual level (Kilduff and Krackhardt, 1994). It is only recently that some researchers have pointed out the importance of individual characteristics (e.g. Mehra et al., 2001) with a subsequent call for more attention tn the subject to “account for the micro-foundations of structural research” (Kilduff and Krackhardt, 1994).

This neglect can be seen as an ironic twist of history because many foundational concepts of social network research came from renowned social psychologists like Lewin. In his field theory Lewin (1936) he modeled human behavior as a function of the person and the environment (=field), which is a basic structural argument relating to the concept of embeddedness. Other influential concepts from social psychologists include Moreno’s soziogram (1934), Heider’s balance theory (1946) and Festinger’s social comparison processes (1954, see Jansen 2003; Kilduff and Krackhardt 1994).

3 Other Research Disciplines

Several fields of research, however, have theoretically treated the question of reciprocal influences between persons and their environment. While the environment is seen as constricting the behavioral options of the person, the individual reciprocally influences his environment guided by his preferences or personality. In psychology, the scientific term “personality” refers to the entire mental organization of a person’s characteristics or traits. A trait is a temporally stable, cross-situational individual attribute. For the past decades, the five-factor model of personality (Big V) has been the widely accepted general taxonomy of personality traits. It states that five major traits, extraversion, neuroticism, agreeableness, conscientiousness and openness to experience, efficiently describe personality differences. To a large extent these traits remain stable throughout the lifetime (Soldz and Vaillant, 1999).

3.1 Social Exchange Theory

Naturally, sociologic theories address the formation and change of social structures. As opposed to social network research, social exchange theory (Blau, 1967; Molm and Cook, 1995) presumes that actors actively pursue their interests. Some of these interests may be satisfied through social interaction, which constitutes the exchange of intrinsically or extrinsically valued rewards (Blau, 1967). It further states that persons act in a self-interested manner in order to maximize their benefits and minimize the costs involved. In contrast to normative economic decision models, social exchange theory seeks to describe and analyze social associations as structures. It is assumed that persons do not necessarily possess complete information, are to some degree restricted by their social environment, and can pursue different goals with shifting preferences (Blau, 1967).

Applying this to social relationships, one can conclude that individuals generally aim to establish relationships that involve maximal returns, e.g. support, entertainment, information, etc., and minimal costs in terms of time, money, humiliation etc. Employing this theoretical background, Klein et al. (2004) investigated whether personality characteristics associated with high benefits or low costs for the advice, friendship or adversarial network would correlate with indegree centrality in the respective network. In line with their reasoning education which promises high informational benefits correlated positively with centrality in the advice network while neuroticism, a Big Five trait associated with high costs, predicted low centrality in the advice as well as friendship networks. Yet, unexpectedly, education also predicted friendship centrality while the Big Five characteristics associated with high benefits like extraversion and openness to experience correlated with centrality in the adversarial network. These results demand further clarifying research.

3.2 Dynamic Interactionism

In the same vein, interactionist theories assume a similar active role on behalf of the person stating that people choose and create their social situation, e.g. their profession, friendships, social activities, etc. As can be seen by the following statement by Lewin (1936), this idea is hardly new : “Every psychological event depends upon the state of the person and at the same time on the environment, although their relative importance is different in different cases.” By virtue of the reference to Lewin, the interactionist paradigm and social network analysis actually share a common root.

Stemming from psychology, interactionism presumes that people select their environment congruent with their dispositions, preferences or attitudes; meanwhile the environment reinforces and shapes the individual and his behavior: “people foster consistent social environments which then reciprocate by fostering behavioral consistency” (Bowers, 1973). The dynamic interaction paradigm sees behavior as shaped by the “reciprocal causal relation between personality and environment” (Ickes et al., 1997, p. 167). Extensive research has shown that the situational choices of persons vary as a function of their dispositional factors, e.g. outgoing (extraverted) persons seek out active as well as social leisure situations, while introverts prefer more passive and solitary recreations like reading (Furnham, 1981; Diener et al., 1984). Similarly, persons with high need for achievement spend more time in work situations (Diener et al. 1984; for an overview see Ickes et al. 1997). The reciprocal process is illustrated nicely in a study by Roberts et al. (2003) in which personality traits predicted work experiences 8 years later which in turn predicted changes in personality traits. Furthermore, these relations corresponded insofar that the predictive traits for work experiences were the ones affected by them.

Applying the paradigm to an evolving social environment (yet not integrating the social network methods), Asendorpf and Wilpers (1998) surveyed new university students repeatedly during their first 18 months at university. Sociable students established more friendships, interacted more and had more friends while shyness had the opposite effect on building friendships and interaction frequency. Agreeableness and conscientiousness influenced the number of conflicts and contact frequency, respectively. These results lend substantive support to the dynamic interactionist perspective, leading to the assertion that people choose their social relationships just as they choose their environments (Ickes et al., 1997).

3.3 Game Theory and Network Formation

In the field of economics, theories start to address individual differences. Within the rational choice paradigm persons are thought to play an active role by choosing an action alternative which, given their resources, best meets their preferences. However, non-monetary preferences are still difficult to measure and therefore to incorporate into the expected utility model. The behavioral scope of persons is therefore restricted by the assumption that individuals uniformly maximize their expected utility.

However, current studies in the field of game theory show that personal attributes (be they called traits or social motives) influence individual choice behavior in the form that persons do not always maximize their benefits. For example, Berninghaus et al. (2005b) demonstrated in two different experiments the effects of the individual attribute inequity aversion on network formation. In the experiments, subjects could choose to initiate contacts with others and thereby gain information benefits, but had to pay for the contact; meanwhile the receipient of the contact paid nothing, akin to receiving a telephone call. Many subject teams deviated minimally from the optimal communication structures in which one player would have had a higher payoff than the others. Inequity aversion seems to prevent persons from maximizing their own benefits; individuals tend to behave in ways that level the benefits of the group. This finding was even more salient when the inequality between the different positions in the network was more pronounced (Berninghaus et al., 2005a). While this explanation is very appealing because of its plausibility and parsimony, it has yet to corroborate its tentative interpretations with individual empirical data. It therefore seems prudent to measure inequality aversion on an individual basis and correlate it to decisions in the network game.

4 Empirical Evidence

Other research has probed the connection between individual characteristics and social structure on a purely empirical basis. In one of the few early studies employing social network background, similar persons occupied similar structural positions (Breiger and Ennis, 1979). Personality was also found to vary with bridging structural holes, which describes holding relations to otherwise disconnected groups. Persons spanning structural holes are more likely to be independent, open to change and looking for responsibility (Burt et al., 1998).

For social psychology, the field of network attributes gained relevance by virtue of its connection to social support and therefore to important outcomes such as employment and health. In this frame it was shown that extraverts have larger networks (Russel et al., 1997; Sarason et al., 1983). Extraversion also correlates with high contact frequency (Russel et al., 1997) and, along with conscientiousness, contributes to networking intensity (Wanberg et al., 2000). Furthermore, it predicts networking-related constructs such as the number of leadership roles, popularity and dating variety (De Raad, 2000; Paunonen, 2003; Paunonen and Ashton, 2001). Neurotizism is generally associated negatively with social relations (Wanberg et al., 2000; Asendorpf and Wilpers, 1998; Klein et al., 2004).

Only recent investigations consider constructs of both of the two well-studied fields personality psychology and social network analysis. Kanfer and Tanaka (1993), employing a whole network approach with the students of a seminar, discovered two interesting connections: first, all of the Big Five personality dimensions (with the exception of agreeableness) correlated most closely with indegree; with indegree and outdegree defined as those interactions reportey by others and self-reported, respectively. This indicates that personality plays a significant role in how others perceive interactions. Furthermore, kind (agreeable) persons tended to occupy central positions and reported more interacting with others while outgoing (extraverted) and secure (low neurotizism) persons had more people reporting interacting with them.

In a work environment, it was shown that the personality trait self-monitoring influences the structural position in the friendship network as well as the size of the workflow network (Mehra et al., 2001). The concept of self-monitoring addresses individual differences in public self-portrayal. Highly self-monitoring persons who are thought to be very adaptive to others' expectations in a social situation held more central positions in the informal network and also received and distributed work to more people, i.e. had larger workflow networks than low self-monitors who are portrayed as more intent on being "themselves".

In a different approach, McCarty and Green analyzed the personal networks with respect to the different personality traits and found that it is mainly agreeable and conscientious persons who tend to have well-connected networks (McCarty and Green, 2005). Extraversion and openness to experience correlate to a lower extent with the centralization measures suggesting that extraverted and open persons have more diverse networks, i.e. belong to different, otherwise unconnected subnetworks. These preliminary investigations into the relationship between social structure and personality traits demonstrate the relevance of individual aspects for understanding the origins of network positions and structure.

5 Limitations

The studies discussed here all demonstrate a systematic relationship between the personality of an individual and the social world he inhabits not solely attributable to the embeddedness of a person in the social structure. However, the aforementioned studies can only be seen as preliminary.

Lack of connection model

Several studies have taken a relationship between personality and social network indices in the context of their investigation for granted. However, a detailed model of how personality influences or is influenced by the social structure is missing. Yet, in order to test the assumed relationships, theory-derived falsifiable hypotheses should be posited. Up to now, the results obtained from these analyses have been merely exploratory, since no concrete hypotheses have been derived (Kanfer and Tanaka, 1993; Burt et al., 1998).

Theoretical restrictions

Furthermore, the majority of the discussed studies has taken only one side of the story into account having either employed concepts from psychology or social network research. As discussed in Section 2, only a few studies have been done in the social network field (Breiger and Ennis, 1979; Burt et al., 1998)) and the few presented in section 3 (Asendorpf and Wilpers, 1998; Russel et al., 1997; Wanberg et al., 2000) indicate a similar situation in psychology. Only two recent studies bring these areas together (Mehra et al., 2001; McCarty and Green, 2005). This is a clear indication that much can and should be done in the future. The apparent lack of knowledge about the other discipline has led (Kilduff and Tsai, 2003, p. 10) to comment: "...there seems to be a structural hole between those who focus on social networks and those who focus on the attributes of individuals."

Data

A methodological flaw in the studies under review lies with the data collection method utilized. The network data of all discussed studies was gathered using questionnaires, as were the personality data. This means that the results were based solely on subjective data from one single source. Analyzing protocols of modern Internet and telecommunication networks allows the researcher to directly investigate the existing technology-supported communication networks while avoiding subjective bias. Using different data sources enhances the reliability of the investigation and its results.

Inconsistencies

Finally, several contradictory findings call for a more systematic investigation. While, for example, agreeableness correlates positively with network centrality in various studies (Klein et al., 2004; Kanfer and Tanaka, 1993; McCarty and Green, 2005) in the longitudinal study of Asendorpf and Wilpers (1998) no such trend could be found. Another inconsistency concerns the effects of extraversion. Whereas many studies support a positive effect of extraversion on different network variables such as network size, indegree and degree centrality (Asendorpf and Wilpers, 1998; Kanfer and Tanaka, 1993; McCarty and Green, 2005), it was also shown to have a negative influence as it correlated with centrality in the adversarial network (Klein et al., 2004). These inconsistencies may be attributable to the study design or to a lack of data reliability which can arise when using one-question measures as Kanfer and Tanaka (1993) did.

The discussed downsides and inconsistencies of the empirical research conducted thus far lead to the conclusion that individual contributions to the structure of our social worlds is as yet a blank spot on the map of social network research.

6 Conclusion

Many roads have yet to be explored, such as causal models explaining differing network positions. Is it possible, as some authors suggest (McCarty and Green, 2005), that the interconnection between personality and social networks differs whether one considers ego-networks incorporating various types of social relations or singular mode whole networks, e.g. all students of a seminar? And will this hypothetical difference vanish if we study complete, multi-relation networks of the type emerging nowadays in the internet? Investigating online networks brings one yet to another field namely, considering the specialities of Internet and computer-mediated communication, both of which influence network formation as well as personality expression.

The reciprocal connection between personality and the emergence of social networks cannot be repudiated any longer. What is needed for a thorough investigation is the groundwork of a sound theory, the integration of the advanced methods of personality psychology and social network research, and a combination of different data sources.

References

Asendorpf, J. and S. Wilpers (1998): “Personality Effects on Social Relationships,” *Journal of Personality and Social Psychology*, 74(6), pp. 1531–1544.

Berninghaus, S., K.-M. Ehrhardt, and M. Ott (2005a): “A Network Experiment in Continuous Time: The Influence of Link Costs,” Discussion Paper on the SFB 405, nr. 05-02.

Berninghaus, S., K.-M. Ehrhardt, M. Ott, and B. Vogt (2005b): “Searching for “Stars” - Recent Experimental Results on Network Formation,” Discussion Paper of the SFB 405, nr. 04-34.

Blau, P. (1967): *Exchange and Power in Social Life*, John Wiley & Sons, New York.

Borgatti, S. and P. Foster (2003): “The Network Paradigm in Organizational Research: A Review and Typology,” *Journal of Management*, 29(6), pp. 991–1013.

Bowers, K. (1973): “Situationism in Psychology: an Analysis and a Critique,” *Psychological Review*, 80, pp. 307–336.

Brass, D., J. Galaskiewicz, H. Greve, and W. Tsai (2004): “Taking Stock of Networks and Organizations: a Multilevel Perspective Brass, D.J., Galaskiewicz, J., Greve, H.R., & Tsai, W. (2004): Taking Stock of Networks and Organizations: a Multilevel Perspective,” *Academy of Management Journal*, 47(6), pp. 795–817.

Breiger, R. and J. Ennis (1979): "Personae and Social Roles: The Network Structure of Personality Types in Small Groups," *Social Psychology Quarterly*, 42(3), pp. 262–270.

Burt, R., J. Jannotta, and J. Mahoney (1998): "Personality Correlates of Structural Holes," *Social Networks*, 20, pp. 63–87.

Chang, C.-L., D.-Y. Chen, and T.-R. Chuang (2002): "Browsing Newsgroups with A Social Network Analyzer," in: *Sixth International Conference on Information Visualisation (IV'02)*, pp. 750–755.

De Raad, B. (2000): *The Big Five Personality Factors: the Psycholexical Approach to Personality*, Hogrefe & Huber, Göttingen.

Diener, E., R. Larsen, and R. Emmons (1984): "Person x Situation Interaction: Choice of Situations and Congruence Response Models," *Journal of Personality and Social Psychology*, 47(3), pp. 580–592.

Domingos, P. and M. Richardson (2001): "Mining the Network Value of Customers," in: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining.*

Furnham, A. (1981): "Personality and Activity Preference," *British Journal of Social Psychology*, 20, pp. 57–68.

Galaskiewicz, J. and S. Wassermann (1994): *Advances in Social Network Analysis: Research in the Social and Behavioral Sciences*, Sage, Thousand Oakes.

Hartfiel, G. and K.-H. Hillmann (1982): *Wörterbuch der Soziologie*, Kröner, Stuttgart.

Haythornthwaite, C. (2002): "Building Social Networks via Computer Networks: Creating and Sustaining Distributed Learning Communities," in: K. Renninger and W. Shumar (eds.), *Building Virtual Communities: Learning and Change in Cyberspace*, Cambridge University Press, New York, pp. 159–190.

Ickes, W., M. Snyder, and S. Garcia (1997): "Personality Influences on the Choice of Situations," in: R. Hogan, J. Johnson, and S. Briggs (eds.), *Handbook of Personality Psychology*, Academic Presse, San Diego, pp. 165–195.

Jansen, D. (2003): *Einführung in die Netzwerkanalyse*, Leske & Budrich, Opladen.

Kanfer, A. and J. Tanaka (1993): "Unraveling the Web of Personality Judgments: The Influence of Social Networks on Personality Assessment," *Journal of Personality*, 61(4), pp. 711–738.

Kilduff, M. and D. Krackhardt (1994): "Bringing the Individual Back in: a Structural Analysis of the Internal Market for Reputation in Organizations," *Academy of Management Journal*, 37(1), pp. 87–108.

Kilduff, M. and W. Tsai (2003): *Social Networks and Organizations*, Sage, London.

Klein, K., B.-C. Lim, J. Saltz, and D. Mayer (2004): "How do They Get There? An Examinations of the Antecedents of Centrality in Team Networks," *Academy of Management Journal*, 47(6), pp. 952–963.

Lewin, K. (1936): *A Dynamic Theory of Personality*, McGraw-Hill, New York.

McCarty, C. and J. Green, H.D. (2005): "Personality and Personal Networks," in: *Sunbelt XXV - International Sunbelt Social Network Conference, Konferenzbeitrag.*

Mehra, A., M. Kilduff, and D. Brass (2001): "The Social Networks of High and Low Self-Monitors: Implications for Workplace Performance," *Administrative Science Quarterly*, 46, pp. 121–146.

Molm, L. and K. Cook (1995): "Social Exchange and Exchange Networks," in: K. Cook and J. Fine, G.A.and House (eds.), *Sociological Perspectives on Social Psychology*, Allyn & Bacon, Boston, pp. 209–235.

Paunonen, S. (2003): "Big Five Factors of Personality and Replicated Predictions of Behavior," *Journal of Personality and Social Psychology*, 84(2), pp. 411–424.

Paunonen, S. and M. Ashton (2001): "Big Five Factors and Facets and the Prediction of Behavior," *Journal of Personality and Social Psychology*, 81(3), pp. 525–539.

Roberts, B., A. Caspi, and T. Moffit (2003): "Work Experiences and Personality Development in Young Adulthood," *Journal of Personality and Social Psychology*, 84(3), pp. 582–593.

Russel, D., B. Booth, R. D., and P. Laughlin (1997): "Personality, Social Networks, and Perceived Social Support among Alcoholics: A Structural Equation Analysis," *Journal of Personality*, 65(3), pp. 649–692.

Sarason, I., H. Levine, B. R.B., and B. Sarason (1983): "Assessing Social Support: The Social Support Questionnaire," *Journal of Personality*, 65(3), pp. 649–692.

Schweizer, T. (1989): *Netzwerkanalyse: Ethnologische Perspektiven*, Dietrich Reimer Verlag, Berlin.

Soldz, S. and G. Vaillant (1999): "The Big Five Personality Traits and the Life Course: A 45-Year Longitudinal Study," *Journal of Research in Personality*, 33(2), pp. 208–232.

Wanberg, C., R. Kanfer, and J. Banas (2000): "Predictors and Outcomes of Networking Intensity Among Unemployed Job Seekers," *Journal of Applied Psychology*, 85(4), pp. 491–503.

Wassermann, S. and K. Faust (1994): *Social Network Analysis, Methods and Applications*, Cambridge University Press, Cambridge.

ime

Information Management
and Market Engineering

The research program "Information Management and Market Engineering" focuses on the analysis and the design of electronic markets. Taking a holistic view of the conceptualization and realization of solutions, the research integrates the disciplines business administration, economics, computer science, and law. Topics of interest range from the implementation, quality assurance, and further development of electronic markets to their integration into business processes, innovative business models, and legal frameworks.

The papers assembled in this volume represent an overview of first research results of the Graduate School "Information Management and Market Engineering", which was established at the Universität Karlsruhe (TH) in 2004 and is funded by the Deutsche Forschungsgemeinschaft (German Research Foundation).

ISSN: 1862-8893
ISBN-13: 978-3-86644-059-3
ISBN-10: 3-86644-059-6

www.uvka.de