THESIS ON INFORMATICS AND SYSTEM ENGINEERING C94

Methods for Coopetition and Retention Analysis: An Application to University Management

JAAN ÜBI

TUT PRESS

TALLINN UNIVERSITY OF TECHNOLOGY
Faculty of Information Technology
Department of Informatics

Dissertation was accepted for the defense of the degree of Doctor of Philosophy in Computer and Systems Engineering on May 11, 2014

Supervisors: Innar Liiv, Associate Professor,
Department of Informatics, Tallinn University of Technology and
Mati Tombak, Senior Research Scientist,
Department of Informatics, Tallinn University of Technology

Opponents: Prof. Ah Chung Tsoi
Macau University of Science and Technology, China

Prof. James J. Cochran
Louisiana Tech University, USA

Defense of the thesis: June 18, 2014

Declaration:
I hereby declare that this doctoral thesis, my original investigation and achievement, submitted for the doctoral degree at Tallinn University of Technology, has not been submitted for any academic degree.

/Jaan Übi/

European Union
European Social Fund

Investing in your future

ISSN 1406-4731
ISBN 978-9949-23-630-5 (publication)
ISBN 978-9949-23-631-2 (PDF)

INFORMAATIKA JA SÜSTEEMITEHNIKA C94

Koopetitsiooni ja väljalangevuse uurimise meetodid: rakendus ülikooli juhtimises

JAAN ÜBI

TTÜ KIRJASTUS

SUMMARY

The goal of this doctoral thesis is to develop methods for analyzing two university management indicators: coopetition and retention. Coopetition, which takes place between departments *for* students, is an indicator that we seek to develop. Student retention, in accordance with the need to continue the student university relationship until its successful conclusion, is a well-known indicator that we aim to quantify.

As regards the theoretical side, we turn to business theories and draw parallels with educational management. In order to conduct the empirical analysis, our prevailing method is linear programming, which we apply to both coopetition and retention analysis.

Coopetition is a ubiquitous phenomenon at different levels of the society. Multinational corporation management research can be utilized for coopetition measurement in universities. It illustrates that a department's primary responsibility is fulfilling its mandate – to groom its students. Subsequently it is also a center which has to be useful for the rest of the university through the interlinked curricula.

Coopetition is measured on a social network, where communities of students mark departments. We consider an adjacency matrix, wherefrom we remove layers in order to determine how much noise needs to be removed to best detect the communities. We also permute the drilled down subsets of the adjacency matrix, in order to understand the internal order within the detected subclusters using the created tool Visual Matrix Explorer.

We next arrive at the conclusion that it is suitable to use theories of business network relationships and customer churn in parallel with student retention topics in order to prepare for student retention quantification. We classify student retention data by using the deviation variable and Charnes-Cooper transformation ensemble linear programming discriminant analysis that has implemented bootstrapping, feature salience detection and histogram based decision making.

The main contribution of this thesis is in finding that as the business processes are becoming more and more dynamic, networked and multilateral, these can inform the educational actors by providing useful managerial tools – as the central imperative is still to influence the institutional outcome.

KOKKUVÕTE

Selle doktoritöö eesmärgiks on kahe ülikooli juhtimise indikaatori – koopetitsiooni ja väljalangevuse – analüüsimiseks vajalike meetodite väljatöötamine. Koopetitsioon, mis leiab aset instituutide vahel ja üliõpilaste eest, on indikaator, mille loomine on meie esimeseks eesmärgiks. Tudengite väljalangevus on üldtuntud indikaator, mis langeb kokku vajadusega kindlustada, et tudengi ja ülikooli vaheline suhe jätkuks kuni tudengi eduka lõpetamiseni ja selle kvantifitseerimine on meie teiseks eesmärgiks.

Teoreetilise külje pealt pöördume me äriteooriate poole ja toome paralleele hariduse juhtimisega. Empiirilise uurimuse sooritamiseks on meie põhimeetodiks lineaarne planeerimine, mida me kasutame nii koopetitsiooni kui tudengite väljalangevuse uurimiseks.

Koopetitsioon on läbiv fenomen, mis eksisteerib ühiskonna erinevatel tasanditel. Multinatsionaalsete korporatsioonide juhtimise teooriat saab kasutada koopetitsiooni mõõtmiseks ülikoolides. Näitame, et instituudi peamiseks ülesandeks on oma mandaadi täitmine – oma tudengite koolitamine – ja seejärel on see ka keskus, mis peab olema kasulik läbipõimunud õppekavade kaudu kogu ülejäänud ülikoolile.

Koopetitsiooni mõõdetakse sotsiaalvõrgustikul, kus tudengite kogukonnad märgivad instituute. Me vaatleme naabrusmaatriksit, mille kihte eemaldades saame teada, kui palju müra tuleb kustutada, selleks et kogukonnad oleksid kõige paremini leitud. Lisaks permuteerime me naabrusmaatriksi väljavalitud alamosasid, tegemaks kindlaks loodud tööriista Visual Matrix Exploreri abil alamklastrite sisemise struktuuri.

Järgmisena järeldame, et ärivõrgustike suhete- ja klientide lahkumise teooriaid saab kasutada tudengite väljalangevuse uurimisel, selleks et jõuda tudengite väljalangevuse kvantifitseerimiseni. Me klassifitseerime tudengeid, kasutades absoluutväärtuse võtmise ja Charnes-Cooperi teisenduse ansamblimeetodit lineaarse planeerimisega diskriminantanalüüsil, milles on rakendatud bootstrapping, tähtsamate dimensioonide eraldamine ja histogrammidel rajanev otsuse langetamine.

Doktortöö peamiseks tulemuseks on näidata, et äriprotsesside muutumisel järjest dünaamilisemaks, võrgustunumaks ja mitmepoolsemaks, saab neid kasutada hariduskontekstis, loomaks kasulikke juhtimisvahendeid – olukorras, kus keskne süsteem soovib endiselt institutsionaalset väljundit mõjutada.

ACKNOWLEDGMENTS

Most of the things in our lives happen in one way or another because of gravity. We are helped the most by those near to us; are influenced the most by people who are great but also sufficiently near to us; and only once every so often perturbate randomly, by giving a beggar a hundred bucks. I would like to thank my supervisor, Professor Innar Liiv, for fulfilling both requirements – being both great and near, he has influenced my life in a way that, all in all, only three other people have. If any star system has no more than a couple of stars and a handful of planets, when we look at life from a heliocentric perspective, his presence has been an entry to the system that has transcended and transformed it.

I would like to thank all my supervisors, and there are several :) I would like to thank the second supervisor of my doctoral thesis, Mati Tombak, for his good, down to earth advice over the years.

My greatest heartfelt admiration and respect go to Professor *emeritus* Leo Võhandu, whom I have been lucky enough to be able to call for help on a daily basis. And I literally mean on a daily basis. If my first supervisor is sagacious, Leo's insights are simply illuminating, placing him in a sapient class of his own.

My fourth supervisor is the first among my loved ones to be mentioned – my father. As important as the articles we wrote together have been for me, and as much as I am thankful for being able to bounce all these ideas off you, this thank you has a wider significance. You have been the person who won the battle between Arts and Sciences for me at an early age. Well, I suppose I made up my own mind and followed your path :).

My fifth supervisor, completing this list, is the supervisor of my Master's thesis – Jüri Vilipõld. Also, it's thanks to him that I am working in our department.

Ten years ago, almost exactly to the date, I started working at Tallinn University of Technology. I am very grateful to this wonderful institution for all the opportunities that have come my way, and I hope I have given my best in return as well. Looking back at the last thirty years, my life seems to go in ten year cycles, so it is nice to be right on time with this defense. I am thankful to all my wonderful colleagues, but especially so to our Departmental Chair, Rein Kuusik, who is the heart and soul of this place.

Conversations with my neighbors on plane flights seem to be the beginning of many things in my life, for instance my upcoming Fulbright placement. One time I had the greatest debate about coopetition with my neighbor, a British gentleman (whose contacts I have sadly lost), on a flight back from Beijing. Both of us came up with a number of approaches, and I have benefitted a lot from this talk, so thank you!

I am thankful to all the people who have pushed me towards concentrating my work on a narrow set of issues and wrapping up my doctoral thesis, especially to Victor Olman and Richard Hoffman.

I am thankful to the editor of this doctoral thesis, Aylin Gayibli, who made the text much better.

I am very grateful to, in an alphabetically permuted order: Elena, Heikki, Kiira, Leelo, Marika, Sven. :)

And, most importantly, I am grateful to my loved ones – others being my mother and brother. We have been lucky to grow up in a family where education was always of the utmost value; where we were loved, which provided us with all the resources we have ever needed and has always set the highest goals. The big question in our home was Arts or Science? Although I have chosen the dismal science only, and was also definitely wide-eyed and innocent, I have been guided towards setting rectified goals in life and striving forward, which as time goes on, I learn to value more and more.

I personally feel that coopetition is ubiquitous, for I see it wherever I look in my life. Therefore, last but certainly not least I would like to thank my relatives and friends – my coopetition partners in life – for always being there for me.

Jaan Übi,
Tallinn, 2014

TABLE OF CONTENTS

LIST OF PUBLICATIONS

All publications are reprinted in the appendices of the thesis.

A. Übi, J.; Liiv, I.; Übi, E.; Võhandu, L. (2013). An analysis of community structure detection for educational coopetition. The 2nd IEEE International Conference on E-Learning and E-Technologies in Education (ICEEE2013), Lodz, Poland, September 23-25, 2013. IEEE, 2013, 104 – 109

B. Übi, J.; Übi, E.; Liiv, I.; Võhandu, L. (2013). Predicting student retention by comparing histograms of bootstrapping for Charnes-Cooper transformation-linear programming discriminant analysis. The 2nd IEEE International Conference on E-Learning and E-Technologies in Education (ICEEE2013), Lodz, Poland, September 23-25, 2013. IEEE, 2013, 110 – 114

C. Liiv, I.; Öpik, R.; Übi, J.; Stasko, J. (2012). Visual matrix explorer for collaborative seriation. Wiley Interdisciplinary Reviews: Computational Statistics, 4(1), 85 – 97

D. Übi, J.; Liiv, I. (2010). A Review of Student Churn in the Light of Theories on Business Relationships. The Third International Conference on Educational Data Mining, EDM2010. Pittsburg, USA:, 2010, 329 – 330

AUTHOR'S CONTRIBUTION TO THE PUBLICATIONS

Publication A Author devised the algorithm, undertook computations, applied these in a domain specific way and was responsible for writing the publication. **Section 3 is based on this publication.**

Publication B Author devised the algorithm, undertook computations and was responsible for writing the publication. **Section 5 is based on this publication.**

Publication C Author was responsible for the application of the collaborative seriation method that the article devises and for writing the publication. The underlying software was developed by a co-author. **Section 2 is based on this publication.**

Publication D Author was responsible for developing the theory and for writing the publication. **Section 4 is based on this publication.**

Section 1 is an original contribution of this thesis.

LIST OF ABBREVIATIONS AND DEFINITIONS

MNC	Multinational Corporation
WPM	World Product Mandate
COE	Center of Excellence
VME	Visual Matrix Explorer
BEA	Bond Energy Algorithm
LP	Linear Programming
ALT	Arm’s length transactions
HF	Hierarchical fiat
MCLP	Multiple criteria linear programming
CCT	Charnes-Cooper transformation
DV	Deviation variables method of multiple criteria linear programming

- **Coopetition** – simultaneous competition and cooperation occurring at different levels of the society
- **Managing student retention** – actively managing the relationship with the students, the efforts and proactive action taken so that the relationship continues until the successful conclusion; also includes the reactive measures taken, but those have a short feedback loop (e.g. include corrections to the accession procedures or tests)
- **Managing student churn –** gaining an understanding of and taking reactive measures based on the students who have dropped out, so that other students do not churn in the future; these have a wider feedback loop (e.g. include corrections to the number of part-time students accepted)

INTRODUCTION

"If you want to run fast, run alone, if you want to run far, run together"
African Proverb

This doctoral thesis is about running far together – it is about coopetition, which is a pervasive phenomenon at various levels of the society. At the beginning, coopetition tends to be something that we come up with when describing the innate nature of a somewhat "fuzzy" process. However, as time goes on, we simply learn to come to terms with it as we recognize yet another coopetitive state of affairs and learn how to benefit from it. This doctoral thesis is about going far, as it is also about student retention, i.e. retaining the student-university relationship all the way to its successful conclusion.

To begin with, looking at coopetition, we can see a situation in cycling that is somewhat akin to the proverb about running (see Figure 1: The coopetitive process untangled). There are a number of teams, each with a designated winner at the finish line, competing for the first place. Cyclists take turns during which the lead-out man does most of the work and has others following him in the slipstream. As the lead-out man tires, he is peeled off and may even be dropped by the peloton (pack), and the next *domestique* of the main sprinter takes the lead. Finally, perhaps even only at the last few hundred meters the main sprinter, with a burst of speed, wins the race. This way the bikers are competing and cooperating at the same time. Men from different teams will in turn be the lead-outs. Thus, we see that the coopetitive process is further complicated as there is **simultaneous coopetition both on personal as well as on the team level**. There have been combinations of cooperation and competition, wherein two weaker cycling teams have cooperated to not let the strongest team win, but as a British idiom goes, "it is not cricket" (it is unfair). Thus, there remains the question of the **fairness of a coopetitive process**, as judged by the antitrust law of the state.

Returning to the personal level, coopetition is evident in universities between doctoral students. On the one hand, they are part of a research group, certainly working towards a common goal, which can factually be observed in the form of joint publications. This is also evident in how theories are versed in common discussions. Yet, at the same time there is competition for better placement in the university after the studies have been completed. Therein arises the question whether coopetition occurs as **simultaneous competition and cooperation, or** whether there are **subsequent bursts of the two.**

On the organizational level departments coopete. Multinational corporations (MNCs) with their national subsidiaries are a perfect venue for coopetition research. Taking MNCs as an example, we will see coopetition implied by a number of phenomena in this thesis; it has also been explicitly studied. We will, however, measure coopetition between the departments in educational institutions. This is the first university management indicator pertaining to the student university relationship that we undertake to quantify.

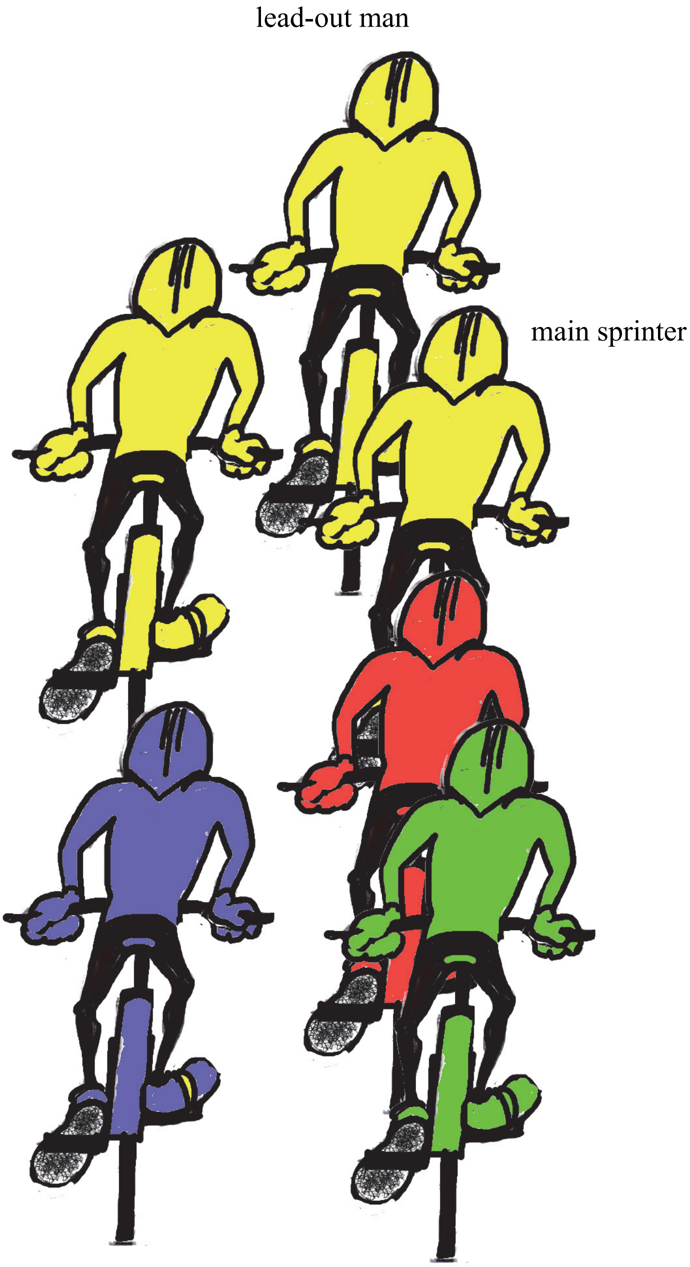


Figure 1 The coopetitive process untangled

Coopetition on the market level has received long scrutiny. A vivid example of this is the big three automakers, who in the processes of competing for customers, sell the car with a coupon for maintenance which is redeemable in the shops of any of the three (Chen, 2008). On the one hand this indicates cooperations – it is as if the maintenance shops are jointly owned. Yet in reality they are not so competition spurs as the shops are exposed to further competition in addition to that inside the given automakers shop network. Thus, the process of coopetition is **often fuzzy,** which in this thesis will give raise to conceptualizations such as the coopetitive duality, wherein competition and cooperation are like light and dark – duals that define each other interdependently. Coopetition at the industrial cluster level puts forth another idea. Coopetition often occurs in the form of two parties being **subjects that coopete *against* an object** – the third party – in a similar manner to how the German and Spanish cyclists coopeted against the superior British team in the 2012 Olympics.

While the cyclists or industrial cluster members coopeted *against* an object, coopetition can occur ***for* an object** as well. The semantic constructs *to be in competition for better students, to be in cooperation for better students* and *to be in coopetition for better students,* have notable differences in the way the preposition *for* is used.

Authors have asked (Harding, 2004) whether on the societal level the general emphasis has been extensively on competition and thus efficiency, at the expense of our social nature, while we must rely more on cooperation. Nevertheless, we agree with the rebuttal (Walker et al., 2004) which remarks that competition and cooperation have both coexisted and evolved in the society.

Discussions about competition and cooperation can be depicted **in a pendulum-like fashion,** as something that swings back and forth from one condition to another. Thus, for instance, intraorganizational competition was idealized when Nokia propounded it. Yet when Nokia failed, the focus in research moved more towards emphasizing "coevolving systems" in organizations. In some situations, coopetition has proven to be the middle ground in this pendulum.

Coming back to the personal level, (Simmons, 2013) has outlined the differences between *efficiency oriented* and *relational thinking,* with the latter much more oriented towards coopetition. A relational thinker will surround himself with people who are assets, not means towards an end, and who go where the system leads them. The article paraphrased the same proverb that we have chosen by stating: "If you want to go fast, go alone, if you want to go far, go together", implying that going far might take you to more meaningful places.

Lastly, let us take the economics level as a whole. Studies using simulation (Yami et al. 2010) have shown that while in the short run (if we just want to run a spurt) competitive action prevails, in the long run (if we want to run far) it is the much more cooperative behaviour that maximizes wellbeing. Therefore, as our goal is to run far together, we must effectively balance competition and cooperation, just as cyclists, relational thinkers, organizations, industrial clusters, companies and the society at large do.

The students of a university, who pass many of the courses together are certainly also in coopetition with each other which leads them right past the graduation date and onto the job market. However, we do not seek to quantify the coopetition of the students, but rather their retention/churn, as we are interested in retaining their student status until the successful completion of their studies. Thus, we consider two university management indicators – we develop a coopetition indicator and help to quantify the well-known student retention indicator.

The study of student retention can take two related directions of research. On the one hand, we can study student retention and the effort that universities make when retaining their students. Therefore, we will identify courses and demographic factors that are important drivers of retention. On the other hand, the focus can be on considering student churn. In that case, churn factors and taxonomy would be of interest. These form the basis for a wider feedback loop which also gives the university a possibility to investigate retrospectively what was not done in the best possible way and to try to right those wrongs. In the case of retention, for example, if accession procedures and criteria can be improved, then churn factors can point us towards a need for general level measures, such as accepting a smaller number of part-time students.

Motivation for the study and research questions

As the area of organizational coopetition has so far been scarcely studied, we choose the multinational corporations that with their geographical, functional, and industrial boundaries have been called archetypal organizations (Buckley and Casson, 1998), in order to draw parallels with universities. Thence originates the idea to draw parallels and **apply business theories** in the educational management and educational data mining context.

We will also see theories, e.g. those dealing with employee turnover (Bean, 1980), that are directly juxtaposed with the student retention topic in order to develop new insights. Student retention appears to be a natural venue for the application of ideas, such as customer lifetime valuewhich, although previously sometimes errorenously applied (e.g. Ackerman and Schibrowsky, 2008), is essential in taking action to prevent retention. This gave us the idea of applying customer churn and business network relationship theories in order to identify relevant factors for measuring student retention.

We may say that on a more general level our goal is to bring parallels between the business and the educational realm. Examples of business management principles being applied in other domains include "the birth of philanthrocapitalism" (Economist, 2006), whereat Bill Gates created his foundation and managed it as a business enterprise. It is important to note that just as parallels can be brought between business management and educational management (as we will see during our investigation of coopetition and retention), there also exist differences. Although, both Google and TUT have defined a mission that they follow ("Google’s mission is to organize the world’s information and make it universally accessible and useful" and "The mission of TUT is to create a synergy between technological, exact, natural, healt and social sciences that promotes the development of the society") the underlying primary purpose of a business enterprise can be stated to be the maximization of shareholder value, whereas a state owned university of technology serves the purpose of providing the society with educated people as well as that of scientific development. Additionally, education is predominantly provided free of charge in Estonia whereas the Ministry of Education and Research is paying for the credits that the students take. However, it is the general managerial principle that we seek to draw parallels with by arguing that educational management can be informed by the businesslike way of thinking. We thereby try to accommodate the differences by, for example, taking into account the academic freedom that is being practiced in universities. We simply consider university to be a special case of organization, as there are also multinational corporations practicing intrapreneurship (Ghoshal and Bartlett, 1997), which are in this thesis shown to be similar to academic freedom.

Our study of university management indicators has to be theoretically scrutinized as well as empirically quantified. As our focus is on using business theories for investigating both coopetition and retention, we use linear programming (LP) for quantifying both. In the case of coopetition, that is coupled with heuristic methods when performing social network analysis. As regards retention, we seek to create an ensemble bootstrapping method that combines two LP approaches. There are a few natural ways of studying coopetition, amongst which simulation studies (e.g. Yami, 2010) and network science (e.g. Hao et al., 2010) stand out. Our main device is applying social network analysis (SNA) and seriation of adjacency matrices. On the one hand, cooperation is well evident from the links which form network communities but so is competition, depending on how we measure the links. **Furthermore, we can see facets of competition from the "missing links" which are factored in while seriating the adjacency matrices of a social network.** Our idea of applying linear programming for discriminant analysis stems form the algorithmic improvement we seek to make – the use of Data Envelopment Analysis and Charnes-Cooper transformation has been suggested but not thoroughly applied in previous literature. We aim to further enchance this approach by using bootstrapping and also by creating an ensemble method.

Hence, the research question of this thesis is (see also Figure 4):

Methods for coopetition and retention analysis: how to apply quantitative methods for developing and predicting various university management indicators, in the light of contemporary management theories?

The general research question has two subquestions that are going to be answered throughout the thesis:

How to measure departmental coopetition for students, in order to manage it effectively? (See Figure 2)

How to predict student retention for managing the student relationship effectively? (See Figure 3)

Figure 2 describes our quest for measuring coopetition. Business theories form an overarching theoretical dome that covers the concepts (competition and cooperation joined into coopetition) of the indicator we are developing. We are planning to develop the indicator by applying social network analysis. The indicator is supported by two pillars-to-be. We are looking forward to investigating how to facilitate measuring the indicator. We are also looking forward to revealing the inner structure of the communities that will be detected. We seek to apply ***the data scientist approach to management*** *in our coopetition study. Therefore we* ***will combine the analysis of the problem with a number of algorithmic improvements, which lead us to an interdisciplinary solution.***

Figure 3 describes our quest for predicting retention. Our plan is to once again use business theories, whereby we identify different factors that are critical to our prediction process. We will develop the retention indicator and identify the input dimensions of its prediction. Finally, we will make a prediction by using linear programming discriminant analysis.

There are a number of minor research questions that warrant an answer (as is evident from Figures 2 and 3), and are given a specific answer in the concluding discussion (developed throughout the thesis):

What parallels can be drawn between organizational coopetition in multinational corporations and in universities? (Answered in Section 1)

What parallels can be drawn between customer churn and the theories of business network relationships on the one hand, and the student-university relationship theories (educational management) on the other hand? (Answered in Section 4)

What methods are useful for developing a coopetition indicator at a university? (Answered in Section 1)

What means of retaining the signal and discarding the noise can be deployed in order to better measure the coopetition indicator on the student similarity/adjacency matrix? (Answered in Section 3)

What can the permutations of the student similarity/adjacency matrix tell us about the communities that are formed? (Answered in Section 2)

What methods can be used for predicting student retention? (Answered in Section 5)

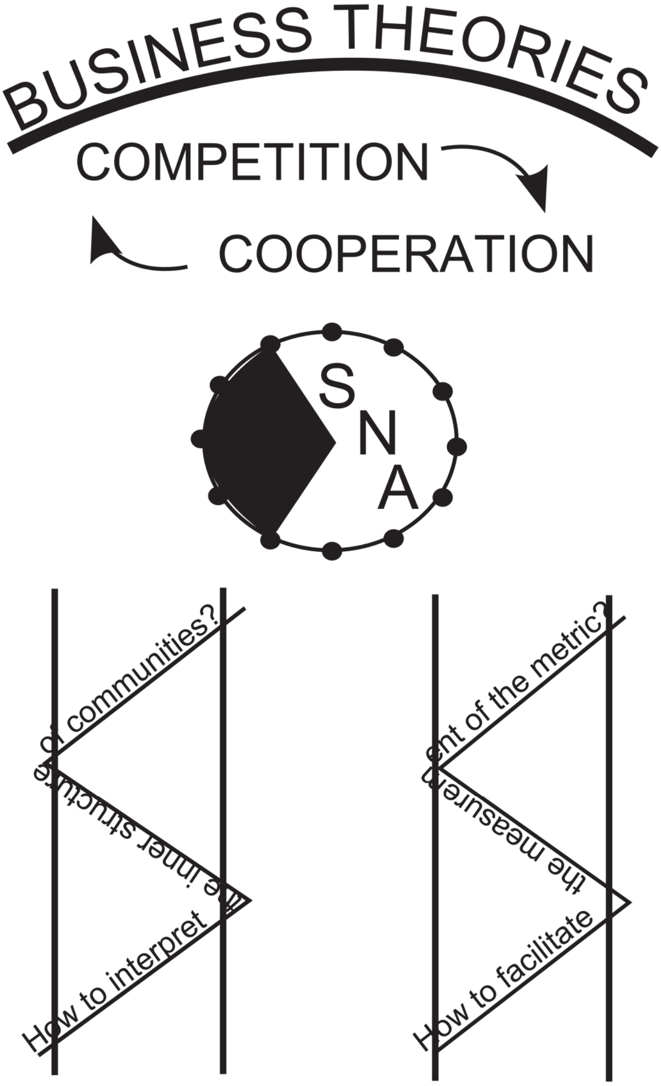

BUSINESS THEORIES
COMPETITION
COOPERATION
SNA
How to interpret the inner structure of communities?
How to facilitate the measurement of the metric?

Figure 2 Setting the scene for the coopetition study

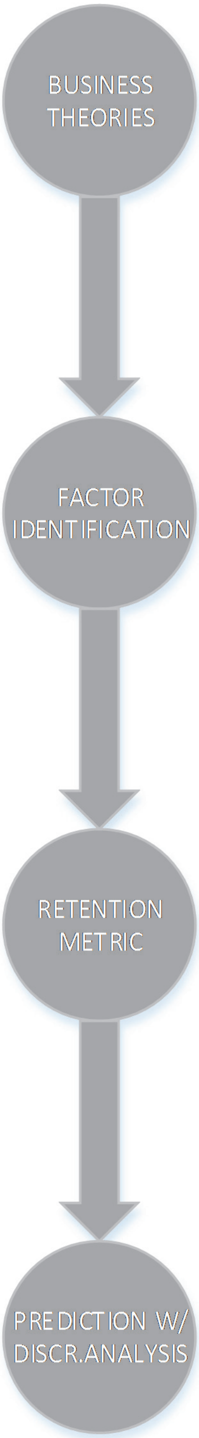


Figure 3 Setting the scene for the retention study

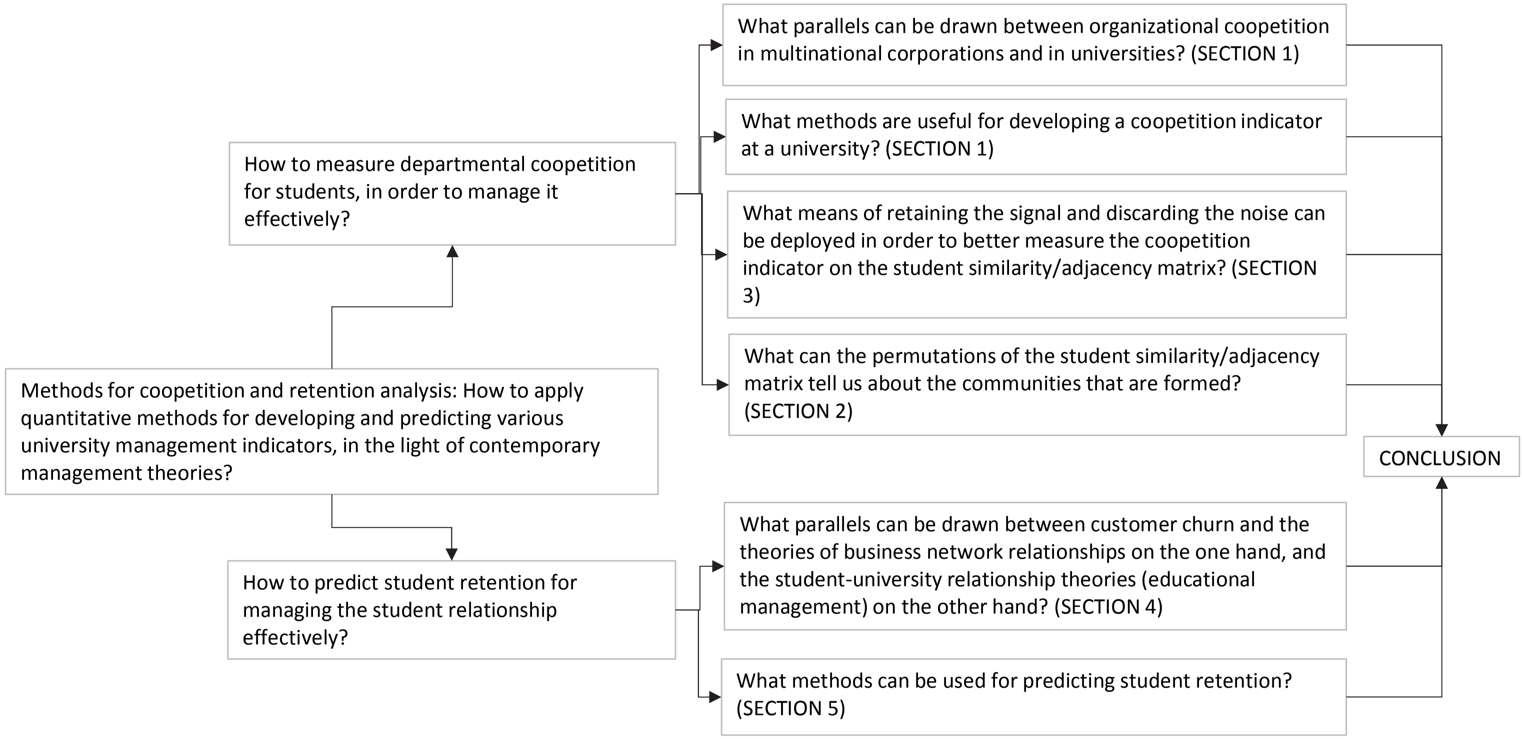

Methods for coopetition and retention analysis: How to apply quantitative methods for developing and predicting various university management indicators, in the light of contemporary management theories?
How to measure departmental coopetition for students, in order to manage it effectively?
How to predict student retention for managing the student relationship effectively?
What parallels can be drawn between organizational coopetition in multinational corporations and in universities? (SECTION 1)
What methods are useful for developing a coopetition indicator at a university? (SECTION 1)
What means of retaining the signal and discarding the noise can be deployed in order to better measure the coopetition indicator on the student similarity/adjacency matrix? (SECTION 3)
What can the permutations of the student similarity/adjacency matrix tell us about the communities that are formed? (SECTION 2)
What parallels can be drawn between customer churn and the theories of business network relationships on the one hand, and the student-university relationship theories (educational management) on the other hand? (SECTION 4)
What methods can be used for predicting student retention? (SECTION 5)
CONCLUSION

Figure 4 Outline of the thesis

1 MEASURING MULTINATIONAL CORPORATION-LIKE COOPETITION IN A UNIVERSITY CONTEXT

The goal of this section is to develop an indicator that quantifies the simultaneous competition and cooperation that takes place in organizations. As the concepts seem to be dichotomous opposites at first, the term internal coopetition duality is put forth. Parallels are drawn between the coopetitive processes in big multinational corporations (MNCs) and those taking place in universities. The structural solutions used in both are also analyzed.

Data mining is used when looking at how specializations inside a university are in coopetition for "better" students. We look at the profiles that students have and find natural divisions between the specializations by applying graph theory and modularity algorithms for community detection. The competitive position of the specializations is evident in the average grades of the detected communities. The ratio of intercommunity ties to intracommunity ties (conductance) quantifies the cooperative stance, however, as students with similar profiles express linkages in the curricula.

We develop an indicator that takes into account both the competitive position and cooperative stance. Taking care of the management of coopetition duality ensures that a department simultaneously achieves its primary goal (grooming its own students) as well as its secondary goal (having a curricula with breadth, that is useful for others), thus maximizing the performance of the entire organization.

1.1 An introduction to a study of coopetition

A decade and a half ago, the term coopetition was coined (Brandenburger and Nalebuff, 1998), indicating the simultaneous competition and cooperation that takes place on the markets. By now there have been calls for (Walley, 2007) and some efforts of studying this on other levels – e.g. society, intrafirm, personal (Ritala et al., 2009; Walker et al., 2004). However, there issome confusion as to how the simultaneous competition and cooperation processes can actually pan out. Our research considers the coopetitive state of affairs within universities. A facet of our research is our conviction that a wide range of business theories, especially those dealing with huge multinational corporations (MNCs), can be used to inform us about processes taking place in educational institutions. Such companies form network organizations with natural national boundaries and strong implications for competition-cooperation.

Whilst performance indicators are definitely useful for choosing an educational institution (Allen and Burgess, 2013), they are important for us for managerial purposes, as we seek to make such an assessment (e.g. Soteriou et al., 1998). Our goal is to empirically quantify the coopetition taking place – this way we can create the means for actively managing the coopetitive process. We are going to apply network science, the importance of which in economic research has risen into focus after the financial crisis (Scweitzer et al., 2009). More specifically, we adopt an organizational analysis point of view (looking at the performance of the departments of a university – see e.g. Messner and Ruhl (1998)) and perform organizational network analysis by evaluating the similarities of student profiles.

This section is organized in the following manner: subsection 1.2 analyzes the relationship between competition and cooperation; subsection 1.3 discussesdifferent conceptualizations of the process on a societal level; subsection 1.4 draws parallels between the situation in MNCs and universities; subsection 1.5 analyzes the structural solutions used that are relevant for our analysis; subsection 1.6 puts forward the empirical investigation of the process at our university and the managerial implications; subsection 1.7 gives conclusive remarks.

1.2 The coopetition duality

In a dichotomy, a whole is split into two non-overlapping parts. In a classic article on "concept misformation" Sartori (1970) argues that concept formation is inherently based on classification and that dichotomies are exclusively fundamental to reasoning about concepts. However, a large body of research in linguistics, cognitive psychology and cognitive science is supporting a more multifaceted view of human cognition, according to which the remarkable capacity of the mind to conceptualize different modes of gradation and different forms of the partial occurrence of phenomena is equally important (Collier and Adcock, 1999). In linguistics, for example, cline is a scale of continuous gradation. Therefore, as both verbs see and kill are transitive (as opposed to intransitive), see is described as having lower transitivity.

When dichotomous concepts at the ends of the continuum are each other's opposites, a paradox is formed (Poole and Van de Ven, 1989), marking the seeming impossibility of seamlessly integrating the two. An example here would be the objects of our research - the processes of competition and cooperation. Here it is difficult to conceptualize a 70% competitive and 30% cooperative arrangement. The paradox of simultaneous competition and cooperation is termed coopetition in business literature (Branderburger and Nalebuff, 1998). Yet just as within and between organizations, we can also see this taking place between individuals in a society, and elsewhere in nature, e.g.with bacteria (Griffin et al., 2004).

Chen (2008) integrates the Western term paradox and the Eastern "middle-way" thinking into a concept of "transparadox" that in the case of coopetition uses three levels– independent, interrelated and interdependent opposites. An example of independent opposites would be the strict choice between competition and cooperation that oligopoly market theory presents for neoclassical economics (Scherer and Ross, 1990). As for interrelated opposites, an example would be a US car company offering a $1000 rebate on car parts, redeemable in an outlet of any market participant (Chen, 2008). We are the most concerned with the third option, the interdependent opposites. It is akin to the relationship between light and dark – one is defined through the other, the two do not have an independent meaning.

We borrow the term "duality" from Chen (2008) to mark what is meant by the third kind of simultaneous competition and cooperation that also takes place within organizations. Duality is widely used in mathematics, for instance in operations research, where a problem and its dual are solved in an interdependent fashion and the values of their maximum/minimum are the same (e.g. Bazaraa et al., 1990).

Thus, we study the internal coopetition duality. While coopetition has been the focus of research for more than a decade and a half by now (e.g. Peng and Bourne, 2009; Ritala, 2011; see Peng et al., 2011 for another application of the Eastern way of thinking), a need for studying internal coopetition has only been stated recently (Walley, 2007). There have only been a few studies, e.g. Ritala et al. (2009) that have considered the link of internal coopetition with knowledge transfer and innovation, as well as a narrowly published earlier effort (Ubi, 2003).

1.3 Coopetition duality in society

As we will see in the next two subsections, there are a number of bodies of work, which consider competition and cooperation as dichotomies – "independent opposites" – in situations where they actually are not, and where thus the recognition of the duality of coopetition would be of benefit.

We start on the societal level with a strand of research advocating the "feminist position" (e.g. Harding, 2004). Informed by neuropsychiatrist research on the difference between women and men (e.g. Brizendine, 2006), it states that male economists have asked questions and drawn conclusions only in a certain way and have not incorporated the "women's way of knowing" (Ferber and Nelson, 1993). It posits that the masculine scientific mainstream has overly emphasized competition, which has its negative connotations, whereas cooperation should be at the helm instead.

The fact that this strand of research draws parallels with the masculinity and femininity dimension in a society is noteworthy. According to the studies assessing national cultures (Hofstede, 2005), the masculinity-femininity dimension is, the only cultural dimension that differentiates countries even after we take into account its wealth, thus coloring our understanding of coopetition duality in international organizations and organizations internationally.

This work sides with the critics of the feminist position (c.f. Walker et al., 2004), who speak of - simultaneous competition and cooperation at the societal level. The basic way of reasoning is as follows – in order to compete there has to be an agreement upon the rules of the process (cooperation) in the first place. Also, as is evident from the division of labor, rational humans learned that cooperative action is more efficient than isolated action. It has become evident throughout the 20th century that cooperation without competition will lead to stagnation as competition is a discovery procedure, providing us with the signals from the market (Walker et al., 2004). Neither component of this duality, competition nor cooperation, can be stated to be the sole "rectified" final goal for humans. In the case of cooperation, our final goal would stem from the social nature of humans, and in the case of competition, it would be to solely increase the efficiency and material progress.

1.4 Coopetition duality in organizations – parallels between multinational corporations and universities

The basic premise of our research into the coopetition duality is the comparison of the state of affairs in educational organizations with those prevailing in commercial organizations, more specifically in multinational corporations (MNCs).

In MNCs the basic units operate in different countries and are called subsidiaries. In the university under consideration the basic units of organization are the departments, and there is as strong a separation between these as there is between the subsidiaries of an MNC. If a faculty member were to move from one department to another, for the management it would be as if they had left the organization altogether.

Humes (1993) discusses managers on the fast track and thus creating carriers of MNC corporate culture. MNCs are essentially three- dimensional (Ubi, 2003) as they have people working at different functions of the company (R&D, marketing, production, etc.); on different products/in different product divisions; and in different countries. Corporate culture is an important tool for MNCs (Hedlund, 1986). According to Humes (1993) different bonds will have to be broken while creating its carriers. For better carrying of corporate culture an employee would have to be transferred between different functions (to break the professional alignment); change their position on the product dimension; and work internationally. As was mentioned before, transfers between the departments of a university are rare. There is also less chance for it becauseoften the content of the work differs fundamentally. This forms a reason for there to be less cooperation than in MNCs.Although the organization only has one basic separating dimension, namely the departmental, there are no classically footloose employees, unlike managers on the fast track for MNCs.

On the other hand, due to each MNC subsidiary being embedded in its local national culture, alignment with it is an additional factor influencing the coopetition duality.Employees who are not expatriates identify mainly with their home country (subsidiary), just like the ordinary faculty members in a university department.

Let us now consider some structural solutions that MNCs use for dealing with simultaneous competition and cooperation. MNC subsidiaries may have developed into World Product Mandates (WPM) (Roth and Morrison 1992). This means that it has acquired control over the full value-added scope (logistics, R&D, production and marketing) of a specific product or product line along with the responsibility of producing for the world market and controlling the entire value chain. On the other hand, MNC subsidiaries may also be so successful in creating certain product divisions, e.g. marketing campaigns, that they are formally acknowledged as a marketing Center-of-Excellence for the Western hemisphere

(COE) (Moore and Birkinshaw 1998). Figure 5 depicts these two structural embodiments (in red and blue) on a subset of a three-dimensional map of an MNC.

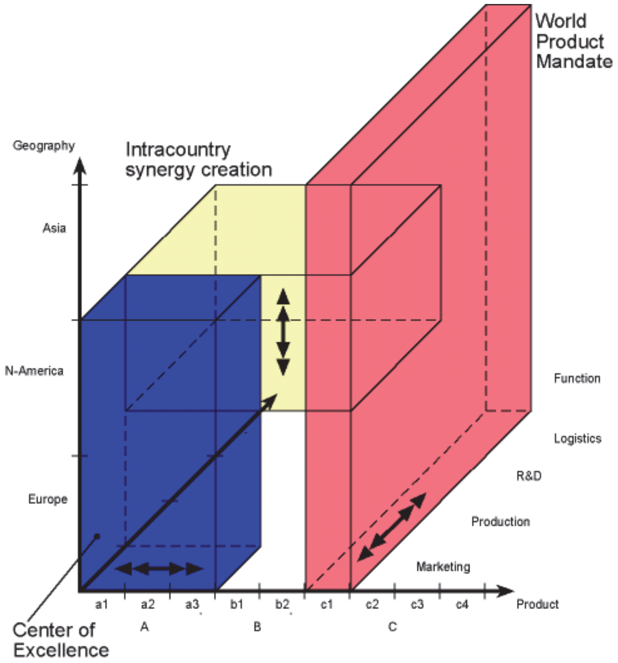


Figure 5 A divergent possibly overlapping structural map of an MNC

We can state that WPM is something that manifests competition within an MNC. It is the result of the development of a subsidiary (Kirstuks, 1999), whereby its capabilities are enhanced. This is best done by taking entrepreneurial action (e.g. Birkinshaw, 2000). For that we may use internal corporate venturing – the creation of divisions with a specific mission for innovation (Burgelman, 1983). Perhaps more relevant for the situation within universities would be using "intrapreneurship" (Ghoshal and Bartlett, 1997), wherein all members of an organization are expected to act on emerging opportunities. According to Birkinshaw (1997), entrepreneurial action is taken in the form of subsidiary initiatives, which can either be directed towards the external marketplace or internally. In the context of universities, external initiatives would display the development of a department's capabilities whereas internal initiatives – winning something, e.g. better performing students, from a competing department – are direct means of internal competition.

It has been widely documented that MNC subsidiaries engage in internal competition (e.g. Krajewski et al., 1994; Galunic and Eisenhardt, 1996) by out-competing a sister unit for an activity, either already performed or upcoming. When the need for internal competition is deemed necessary to be emphasized, terms such as MNCs operating on internal markets are used (c.f. Birkinshaw, 2000; but also similarly Buckley and Casson, 1998). Increased competition is said to be evident from the broad use of internal benchmarking and performance league tables, or from the operations of internal investment agencies.

In a university, another form of internal competition takes place in the budgeting process (c.f. in business organizations Walley, 2007) wherein the amount of money that the state allocates by paying for the tuition of the better performing students is planned for.

At the same time it is important to recognize the need for cooperation inside a business organization like an MNC considering that all subsidiaries are still part of the same company. In management literature there is often a great emphasis on achieving widespread sharing and trust. When leading edge solutions arise in the process of developing subsidiary capabilities, it becomes a natural objective to disseminate these throughout the corporation (c.f. Andersson and Holmström, 2000).

In MNCs' literature on knowledge transfer and organizational learning the managers on a fast track are once again mentioned, as is the development of leading edge IT solutions. The organizational culture needed for such a process is characterized as an open one, based on fairness and shared values.

In the service industry, consulting companies, such as McKinsey, are a fine example; while in the context of industrial firms IKEA stands out (Ghoshal and Bartlett, 1997; also see Heldlund and Ridderstrale, 1995 on projects of international cooperation).

Structurally, we will define COE as the manifestation of such best practice transferal, another example being 3M Sweden, which had leading capabilities in customer-focused marketing and key account management. It was recognized for its strengths and actively helped out other units (Birkinshaw, 2000), being an exemplary knowledge disseminator in an MNC.

In a university cooperation of this kind is evident in doctoral schools that are jointly undertaken between the departments operating in the same field. Departments take part in joint supervisions of interdisciplinary theses. There can be as many as three supervisors, with industry representatives participating as well. There are also horizontal committees that discuss topics like curriculum development, student dropout reduction, industry cooperation, internationalization, etc.

The curricula in our university must contain interlinked parts. There is a minimum number of credits that have to be taken from other departments and faculties. This creates natural linkages between the organizational unitswhich are not merely a manifestation of cooperation. They also bring along a monetary reward as all the declarations have to be paid for. Sometimes there are competing departments, who could provide the same courses. As regards mandates in MNC's, literature has examples of a less than friendly atmosphere that might arise after a subsidiary out-competes a sister for the WPM production of the latter's "charter" (cf. e.g. Birkinshaw,1995). Similar is true for the center-of-excellence dynamics inside a university. If another department will henceforth be providing the general courses in the changing curricula, monetary consequences ensue and this takes a toll on personal relations.

Let us state for our coopetition duality that world product mandates and centers-of-excellence are symbols of its interdependent opposites; that gaining a **mandate** (the right to be the sole proprietor of certain activities, e.g. educating the students in one's own specialization) and being a **center** (to be recognized as an asset valuable enough to be explicitly distributed, e.g. through other departments accessing one's knowledge in the interlinked curricula) are the primary and secondary goal towards which the organizational unit will strive simultaneously.

As can be seen, sometimes it is competition that is emphasized and sometimes cooperation. An example of the latter is the work of Eisenhardt and Galunic (2000). This article defines MNCs as "coevolving systems" and certainly emphasizes the fact that everyone is in the same boat therein.

Still, the two discussions are actually not always isolated. The internal market perspective (Birkinshaw, 2000) has also discussed the phenomenon of knowledge transfer. Corporations are said to have an internal market of capabilities, one that operates without competition and without fees charged for servicing. It is said to be facilitated by strong corporate culture and incentive systems that reward propensity to cooperate. HP and Ericsson are used as examples of companies that capably share while simultaneously competing.

When looking at general accord towards the nature of coopetition, it seems to have evolved over the years in a pendulum-like fashion. For instance, internal competition was much sought after during the time when Nokia propounded as well as implemented it. Since the company got into trouble, the consensus seems to have shifted much more in favor of emphasizing the need for cooperation and sharing.

There are more factors to be examined that cannot be considered in the current work. Studies have looked into different factors that influence the duality of internal coopetition – for instance, (Tsai, 2002) studied the role that the centralization of an MNC has in the internal coopetition between subsidiaries. Another work that considers the dyadic headquarters-subsidiary relationship as well as the network that subsidiaries form is (Ubi, 2003), which finds previous evidence of the role that a subsidiary general manager (SGM) has in the internal coopetition duality. SGM's role was emphasized in both processes that deal with internal competition (a key driver in the internal initiative process) and internal cooperation (information cross-pollinator and creator of horizontal linkages). The implications that this has in the university context – the role of a departmental head in the educational internal coopetition duality and the same role in the departmental centralization – are a subject for further research.

1.5 Internal organization for coopetition – multinational corporations and universities

A central point when considering how MNCs try to find a balance between competition and cooperation is that they should be considered as multicentered/networked, with less emphasis on the overarching hierarchy that might also have been defined. The same is true for universities with their departmentally oriented structure. In the case of MNCs, the term used is heterarchical (Hedlund, 1986; Hedlund and Rolander, 1990; Hedlund, 1993), which points to organizations with a partially overlapping and divergent structural map of Figure 5 but is based on a change of perspectives in a wide range of sciences. The main idea is that the reality is actually organized non-hierarchically and that we are only accustomed to working with it through hierarchies. One good example from complex embodiments is the highest level in evolution – the brain the functioning of which we cannot completely explain but which surely is a non-hierarchical system.

Important contributions are also made by the Uppsala school (e.g. Forsgren and Pedersen, 1998; Forsgren and Pahlberg, 1992) that often does not make a qualitative differentiation between cases where an organizational unit is involved in transactions with sister units and cases where transactions take place with external parties. Some business network articles have also modeled the positions of subsidiaries within MNCs in terms of influence (Forsgren and Pahlberg, 1992; also Ghoshal and Bartlett, 1990) – something that we also aspire to do. Another point propounded by the Uppsala school is that, when considering whether a transaction takes place on the market (at *arm's length*) or through strict *hierarchical fiat*, there also exists a continuum. That is, in business networks we should not consider all the business partners as strictly separable from their relationships. As the arm's length transaction would imply more competition for resources and hierarchical fiat more cooperation thereupon, we find the venue of research to be related with ours. Our empirical analysis is going to view the university as a network of units with a heterarchical structure. Startup companies originating from the university exemplify the fuzzy boundaries that the university has with the industry.

Another point of interest are the works of Burgelman (1983, 1994) that consider autonomous strategic decisions in big corporations. These works show how multiple layers of management are actually involved in taking strategic decisions. This way the managers, who are at first considered to be lower at the organizational drawing board, can through their actions be setting the strategic directions of the company. Intel is an example of this. There the decisions of middle level managers decided the transfer of a memory company into a processor company (Burgelman, 1994). This set of articles also implies a multi-centered view of an organization, and is valid in the context of universities in the sense that there is usually ample space for autonomous strategic initiatives in academia, be it for industry cooperation or research direction choice.

Thus, the profile of the department is developed, and this translates into a competitive position of a mandate to serve the interested students. The development of a departmental profile also influences the position of the department as a center.The courses are necessary for other specializations in the interlinked curricula, which makes the school stronger, even though these inbound interlinkages do not directly help out the department's own graduates.

There is also another body of work that can be drawn in parallel with the discussion on academic freedom, namely the research on “subsidiary slack”. Poynter and White (1985) discuss organizational slack that develops in subsidiaries, and the ways it dissipates. Slack is defined as the excess of total human resources after a proper amount has been allocated for the current strategy. The work shows that subsidiaries have a natural tendency of generating slack and that it can be used for undertaking new value-added activities. As with academic freedom, if initiatives arising from slack are not tolerated, disharmonies arise.

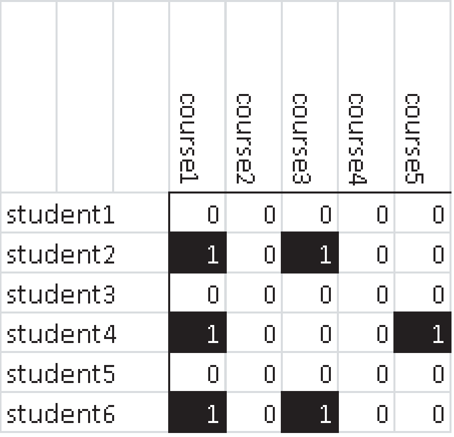
Whether due to autonomous initiatives or slack, the resulting external initiatives, the fruits of which also need to be shared cooperatively, are important for the development of the university.

1.6 Quantifying coopetition in a university

We are using data from the business school of our university. It consists of 509 students, who have graduated from the curriculum “Business” during the time span of 1997-2010. We have information regarding the courses that the students attended, the grades they received, and the final specialization the student chose within the business school. Altogether, students attended 759 courses. The courses were declared 21976 times. The four specializations of the school are: finance, marketing, accounting and management.

As the first step of the analysis we construct a binary matrix (partially displayed on Table 1), with rows representing the students and columns representing the courses attended. This matrix is sparse with 5.66% of cells marked by 1 on the whole. The most frequently taken course was attended 501 times, 463 courses were only attended once, and 71 courses were attended more than 20 times. Thus, the distribution function follows the power law.

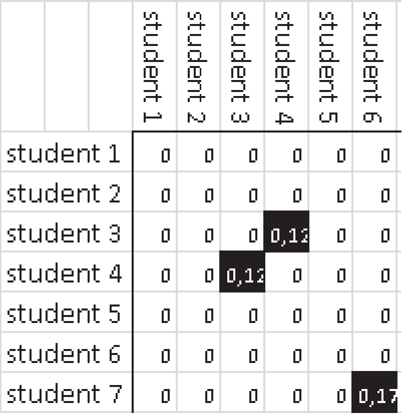
Table 1 A subset of a binary matrix displaying the course selection of students



	course1	course2	course3	course4	course5
student1	0	0	0	0	0
student2	1	0	1	0	0
student3	0	0	0	0	0
student4	1	0	0	0	1
student5	0	0	0	0	0
student6	1	0	1	0	0

We will next construct a 509x509 distance matrix between all the students (see in its final form on Table 2). As the first step we will use Hamming distances, the purpose of which is to count the total number of different course selections by the two students, thus arriving at a distance matrix. For the selection of the distance function, we also tried weighted Hamming distances (the “importance” of a course, the number of times attended, was used as a weight) but this led to significantly poorer community detection, because the course attendance distribution follows the power law and is strongly right skewed. For binary matrices, Hamming distance is equivalent to Manhattan distance and should also be chosen over Euclidean distance as it does not include taking square root in the calculation. We will next calculate the matrix of similarities by taking the reciprocal value of the distance. As the next step we will calculate the histogram of the similarity matrix (on Figure 6), and find out that as the values range from 0..1, 95% of the values are below 0.11. Our final similarity matrix will be composed of those 5% of entries that have value above 0.11. Therefore it is once again sparse, with the same level of sparsity that the original course attendance matrix had (see Table 2). It will only retain information on the most important similarities between students, in parallel with how saturated network studies – e.g. using the reciprocation method – concentrate on the most important linkages (cf. Wejnert, 2010).

Table 2 A subset of a sparse similarity matrix for students that shows the top 5% of similarities for course selection.



	student 1	student 2	student 3	student 4	student 5	student 6
student 1	0	0	0	0	0	0
student 2	0	0	0	0	0	0
student 3	0	0	0	0,12	0	0
student 4	0	0	0,12	0	0	0
student 5	0	0	0	0	0	0
student 6	0	0	0	0	0	0
student 7	0	0	0	0	0	0,17

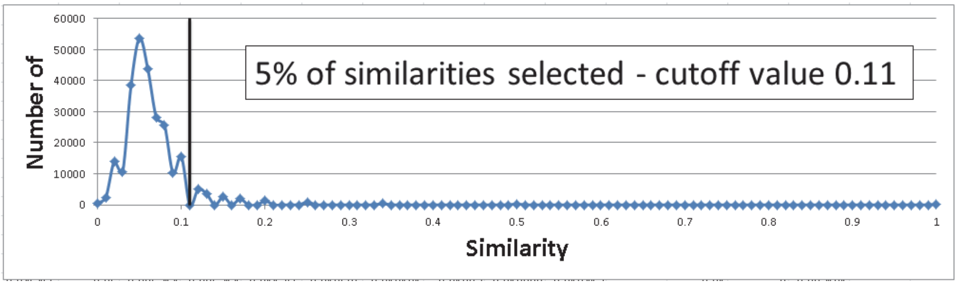


Figure 6 Distribution function of the similarity matrix and the cutoff value.

On Figure 7, we will use ForceAtlas2 algorithm (Jacomy et al., 2011) in order to visualize the similarity matrix/adjacency matrix as a social network. This algorithm has a linear-linear model, with attraction and repulsion forces proportional to the distance between the nodes. The shape of its graph is between Früchterman&Reingold's layout and Noack's LinLog.

As the next step (also on Figure 7) we are going to detect and color the communities. In order to quantify the intercommunity ties, we need an algorithm that does not only find (in sparse graphs) isolated communities with different levels of cliquishness (like Closed Sets algorithms – eg. Lohk et al., 2010; see Hruschka, 2006, for cliquishness), nor only the isolated perfect cliques (like Formal Concept Analysis, e.g. Torim and Lindroos, 2008) but an algorithm that also allows for intercommunity ties. We have chosen the Modularity algorithm (Blondel et al., 2008), which in its class is a comparatively fast performing greedy heuristic (cf. Fortunato, 2010). It builds hierarchical communities in two iterating passes by joining the nodes and building a new structure thereupon. New nodes are joined based on intracommunity linkages, linkages outside the community and the same two kinds of linkages for the focal node being joined.

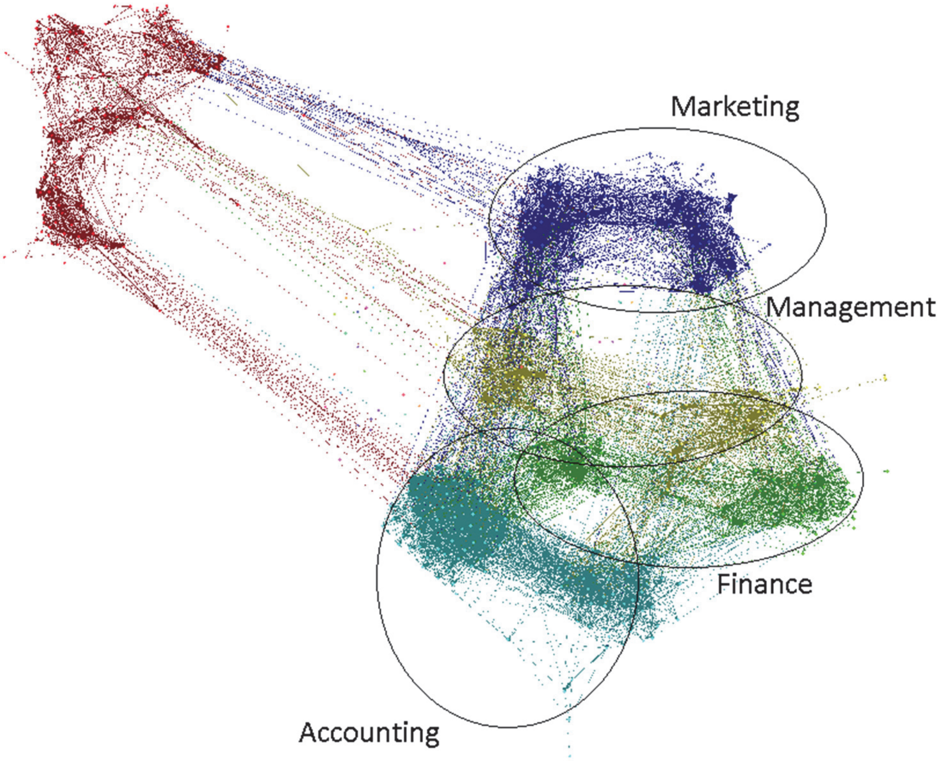


Figure 7 Detection of student specializations for competitive and cooperative stance

As a result, specialization is correctly predicted for 65% of students (331 out of 509). These form the MARKETING (BLUE), MANAGEMENT (YELLOW), FINANCE (GREEN) AND ACCOUNTING (CYAN) communities on the graph.

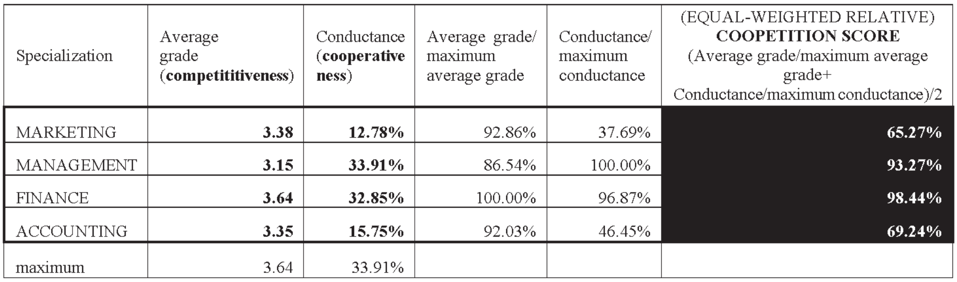
21% of students (108 out of 509) are joined in the unmarked (RED) community, which incorrectly consists of students of all specializations in approximately equal percentages. 13% of students (70 out of 509) are incorrectly not included in any community (these are the solitary dots on the "outside").

In a sense the communities follow the logical structure of the business school: management uses information from marketing and finance; finance is tied to management as its internal consumer but also receives data from accounting.

As is shown on Figure 7, all the correctly detected communities are split into two. The factors accounting for this division are a subject for a follow-up discriminatory analysis. Furthermore, the discriminatory factors accounting for the emergence of the unmarked group should also be studied. On another note, the two sides of the accounting group were also originally split into two, as we would have done by using the discriminant analysis.

Average grades of the groups (on 5 point scale), representing the competitive position of the department mandated with the specialization, are shown on Table 3.

Table 3 Calculation of the coopetition score



Specialization	Average grade (**competititiveness**)	Conductance (**cooperative ness**)	Average grade/ maximum average grade	Conductance/ maximum conductance	(EQUAL-WEIGHTED RELATIVE) **COOPETITION SCORE** (Average grade/maximum average grade+ Conductance/maximum conductance)/2
MARKETING	**3.38**	**12.78%**	92.86%	37.69%	**65.27%**
MANAGEMENT	**3.15**	**33.91%**	86.54%	100.00%	**93.27%**
FINANCE	**3.64**	**32.85%**	100.00%	96.87%	**98.44%**
ACCOUNTING	**3.35**	**15.75%**	92.03%	46.45%	**69.24%**
maximum	3.64	33.91%			

As the strong intragroup linkages define the specializations and allow them to be detected, they express students who have small Hamming distances and correspondingly similar **student profiles**. If the students whom the specialization has been mandated to teach have a high average grade, the specialization is doing competitively well, otherwise it is not.

However, the sparse similarity matrix, with 5% of its values nonzero, also displays linkages between the groups. This means that there are students whose profiles are similar but whose specializations do not coincide. These depict the linkages in curricula, the fact that a student studying, for example, managerial finance, will concentrate both on managerial accounting and on corporate finance. Therefore, this is a student who will constitute a link between the accounting and the finance centers of the graph. As elaborated above, these linkages also display financial ties and the fact that the departments are required to tie together the curricula.

For us the student profiles forming linkages between the departmental specialization centers are **proxies** for the cooperative stance.

For the sparse similarity matrix we consider the following ratio as one expressing the cooperative stance that a specialization has:

$$Cooperation = \frac{\sum all\ linkages\ from\ the\ focal\ specialization\ leading\ to\ others}{\sum all\ intragroup\ linkages\ for\ the\ focal\ specialization} \quad (1)$$

or denoting differently in a social network analysis terms as a traditional conductance (Mislove *et al.*, 2007)

$$K = \frac{\varepsilon_{AB}}{\varepsilon_{AA}} \quad (2)$$

We choose traditional conductance over very similar ones like relative conductance (Mislove et al., 2007) for the ease of interpretation of its results. If the ratio is high, the specialization is doing cooperatively well; otherwise it is not (see Table 3).

As a final step we calculate the (equal-weighted relative) coopetition score. In order to do that, we find the maximums of both the average grade and the conductance. We then calculate how many percentages of the maximum each department achieved – again, for both the average grade and the conductance (see Table 3). The coopetition score is the equal-weighted average of the two previous numbers. It has a range from above zero to 100%. The latter would be achieved if the same department had the highest score for both competition and cooperation for the given time period. The equal-weighted nature of our score stems from the fact that in the swings of the coopetition pendulum discussed above (with the general consensus seemingly changing its preference back and forth between internal competition to cooperation) we find no reason to prefer one over the other.

It can be summarized from the data on the competitive and cooperative positions that finance is the specialization that does very well on both accounts, its overall score being 98.44%. Management, on the other hand, presents a curious case. Its overall score of 93.27% is the second best but it has the worst average grades by far. By this token, we can see that it may pay off to have a curriculum that has breadth and parts of which are useful for different university customers (helping to groom students from competing departments), even though the attraction for the better students is clearly not there.

Managerial implications for university management concern financially rewarding departments for their coopetition score. Depending on the size of the university, this could essentially be a one-off thing happening very rarely; or alternatively could possibly even be undertaken annually on the basis of incoming data on recent graduates. In the context of such big universities, there would be an additional benefit in rewarding the dynamics of coopetition duality. This way the absolute values of the coopetition score play only a partial role, with the recent emerging trends providing additional information.

The action undertaken upon the data does not only have to be restricted to financially rewarding the winners in the ratings. As regards the weaker and at the same time more isolated departments, this provides the university with guiding signals possibly warranting action.

1.7 Conclusive remarks

In the context of the situation inside organizations, this work disagrees with the viewpoint that competition and cooperation form a dichotomous paradox, where either one or the other is present. Instead we side with the call for using the Eastern "middle-way" thinking, which dubs this relationship a duality. In a duality the two interdependent sides cannot exist without, and are defined by, each other –for instance, like light and dark. Considering one might intuitively vouch for there to be more competition on the markets and more cooperation on the intrafirm level, we set out to study the internal coopetition duality.

When considering the studies on a societal level, we see the same question about the nature of the coopetitive relationship raised again. Next we compare the state of affairs in big multinational corporations with those prevailing in universities.

We use the term World Product Mandate (or more generally, mandate) as an exemplifier of the competition that takes place inside MNCs. Meanwhile the Center Of Excellence (again, more generally, center) shows that a national subsidiary of an MNC has been effective in dispersing its unique capabilities. Thus, the difference between being **mandated** and being a **center** throws into sharp relief the nature of the duality of coopetition in an organization. A mandate is achieved either by external or internal initiatives. In the university context, the former would mean the development of departmental capabilities (by research or industry consultation) whereas the latter implies direct competition for better students. Competition also takes place in universities during the budgeting process. While both MNCs and universities have an innate tendency to act parocially, it might be easier for MNCs to be cooperative due to the use of managers on the fast track and the greater internal mobility of workforce in general.

At the same time studies of MNCs underline the widespread sharing and trust that has to exist. One of the studies that emphasizes cooperation considers organizations as coevolving systems. Studies of internal competition have also considered cooperative aspects. In universities, cooperation is evident from doctoral schools, joint supervisions, horizontal commitees and interlinked curricula. Interlinked curricula have both cooperative and competitive aspects.

Organizationally the structure of MNCs has been described as heterarchical, networked and leaving space for autonomous initiatives. The boundaries of the organization are said to be somewhere between the *hierarchical fiat* and *arm's length* transactions. We find the university structure also to be networked and with fuzzy boundaries. Works on MNCs have considered subsidiray slack that is related to autonomous initiatives. Slack is a concept that is also relevant in the university context, because there has to be academic freedom in this networked institution.

The goal of the department is the development of its capabilities, which enables the unit to become better at both its primary goal (grooming its own students) as well as the secondary goal (having a curricula which is useful for

other parts of the school) – the two together maximize the performance of the entire organization.

In order to quantify the coopetitive duality and develop a performance indicator, we consider a binary sparse matrix of the 509 graduates of the business school of our university, and their 759 courses declared. We calculate the reciprocal of Hamming distances between the students and are only concerned with the top 5% of the ties between the students. We use the ForceAtlas2 algorithm for calculating the layout of the adjacency matrix. We next use the Modularity algorithm in order to detect interrelated communities in the social network. As there are four specializations in the business school (marketing, management, finance and accounting), 65% of the students are correctly detected by their specialization using the Modularity algorithm. Further discriminant analysis is warranted in order to delineate the characteristics of the undetected group as well as the binary splits inside the specializations.

The average grade of the detected group makes evident the competitive position of the department inside the business school. At the same time, the ratio of intercommunity ties to intracommunity ties (traditional conductance in a social network) quantifies the cooperative stance. In order to quantify the coopetitive stance of a specialization, we derive a coopetition score as an equal weighted relative indicator of those two. It becomes evident that Finance is the high achiever of the university according to both components; but the curious case of Management, which is the lowest performer as far as the average grade is concerned, also provokes discussion. Namely, it has the highest cooperation score (and correspondingly the second highest total score), underlining the fact that it may pay off to have a curriculum that has breadth and parts of which are useful for different university customers.

Managerial implications for driving the evolution of the dynamics of the coopetition duality include assessing the results, financially rewarding the outcome (possibly repeatedly), and observing the status of low performers; thus making active use of the coopetition score.

Our next goal is to further investigate ways of fine-tuning the coopetition score as well as to deepen our understanding of the social network analysis method used.

2 SERIATION OF COOPETITION MATRICES WITH VISUAL MATRIX EXPLORER

Our goal for this secion is to show, firstly, that coming up with the correct row order of an adjacency matrix is an NP-hard problem that entails finding a block diagonal matrix form and, secondly, to link the discussion to a number of other areas, which by doing seriation work towards similar goals.

We further discuss a collaborative seriation tool (Visual Matrix Explorer, (VME, 2011)) that has been developed, enabling one to compare the results of nine seriation algorithms. VME lets one select subsets of data (to be colored on all respective permutations) and to recursively drill down, having the table rendered again, now with a bigger number of degrees of freedom.

We apply Visual Matrix Explorer to our coopetition matrix from the previous section and perform a number of operations. By the end, we find the internal order within the subdivisions of Marketing, as we perform a drilldown operation on a subset of one rendering and use the BEA algorithm therein. This result overcomes the shortcomings of ordinary clustering, where clusters are formed but the leaves within the clusters remain in an unordered state.

2.1 Permutations of an adjacency matrix

In section 1 we constructed a student similarity matrix, which looked at the reciprocal value of the Hamming disctance that the student-course declaration matrix had. We made the matrix sparser and arrived at an adjacency matrix for our social network. We detected communities that represent the departments by using the modularity algorithm.

In the initial calculation of the adjacency matrix the order of rows is essentially random – it might be determined by the order in which the students are queried from the database.

Finding the correct order of rows/columns is a process that has many similar disguises in related fields of research and has, for example, been termed seriation, matrix reordering, permutation and biclustering (see e.g. (Lenstra, 1974), (Niermann, 2005), (Mueller et al., 2007), (Li et al., 2009)). Seriation, a process that has been defined as "simplifying without destroying" (Bertin, 1981), has been studied since the end of the 19th century. Recently, (Liiv, 2010a) published an historical overview and put forth a unified view. If we, for instance, calculate a correlation matrix for the columns of our data table, information is indeed transformed and some of it lost; however by just permuting everything is retained.

The algorithms that perform seriation are often heuristics that are aiming to increase the "clumpiness" of the permuted matrix through maximizing the various ad-hoc optimization criteria. While we discuss the details of modularity in section 3, other examples include a greedy heuristic BEA (McCormick et al., 1972) that rebuilds the matrix around the more substantial chunks and the ClustanGraphics (Wishart, 1999), which acts on the information on distances that rows or columns have. Some optimization criteria even try to link the discussion of finding the best possible order in the matrix to Kolmogorov complexity and the shortest possible description (Liiv 2010b).

Let us put forth the first example of the problem we are trying to tackle, this time figuratively. Let us go back to Figure 1 (in the introduction, on page 15) that has a clear-cut depiction of coopetition in cycling. We may look at it as at a bitmap that has 6 colors. The initial bitmap contains information in the form of a drawing, as according to its heading it untangles coopetition for us. This 6-color matrix has as many different possible values; it is comprised of 4960 rows and 3508 columns. Such a data table has n!*m!=4960!*3508! possible permutations that the picture may take – all the while no other information is being destroyed, besides the fact that we have changed the row/column order.

However, Figure 8 (The coopetitive process tangled) has one example of this process of "reorganizing without destroying", which effectively **depicts noise**. Stating the obvious – only a minuscule amount of those permutations (perhaps only one) carries any meaningful information – allows us to appreciate the difference in the amount of "energy" spent in trying to assemble this picture from its noisy state as compared to just mixing up the order of rows and columns. The order, seemingly ever so elusive, **must** be hard to find. The relationship between Figure 1 and Figure 8 also pertains to the discussion of information encoding and decoding, whereas the latter will take a lot of computational effort.

We may essentially be dealing with either two-mode data tables (NxM) or those that are two-way one-mode (NxN) (Caroll and Arabie, 1983). Although the number of possible permutations is smaller in the case of a two-way one-mode matrix (n! instead of n!*m!), both problems are still NP-hard by nature. As we will see in detail in section 3, we may either choose to use approximation algorithms in the form of heuristics, or only solve very small problems by finding the strictly optimal solution.

Having moved from the final state in the matrix (Figure 1) to the supposedly initial one (Figure 8), and contemplated the complexity of performing the reversal, we can now turn to our work on Visual Matrix Explorer in search for the right permutation. The developed Visual Matrix Explorer tool enables cross-disciplinary theory brushing by applying nine seriation methods (all with different objective functions) and visualizing the results in a dynamic way (see Figure 12). Therefore, after an interesting permutation has been identified, one can further select a subregion of the matrix, drill down and perform the seriation again. This time we are dealing with a region that has parts of rows and parts of columns cut off, so we have increased the number of degrees of freedom in performing our permutations. Furthermore, all nine algorithms will be reapplied for this subset.

After we identify an interesting chunk in data, it is colored and selected on all the successive plots of the current drilldown level, intuitively showing where this chunk lies on different permutations. If we find a block appealing, we can now appease our interest and check whether it is contiguous according to "other theories" as well.

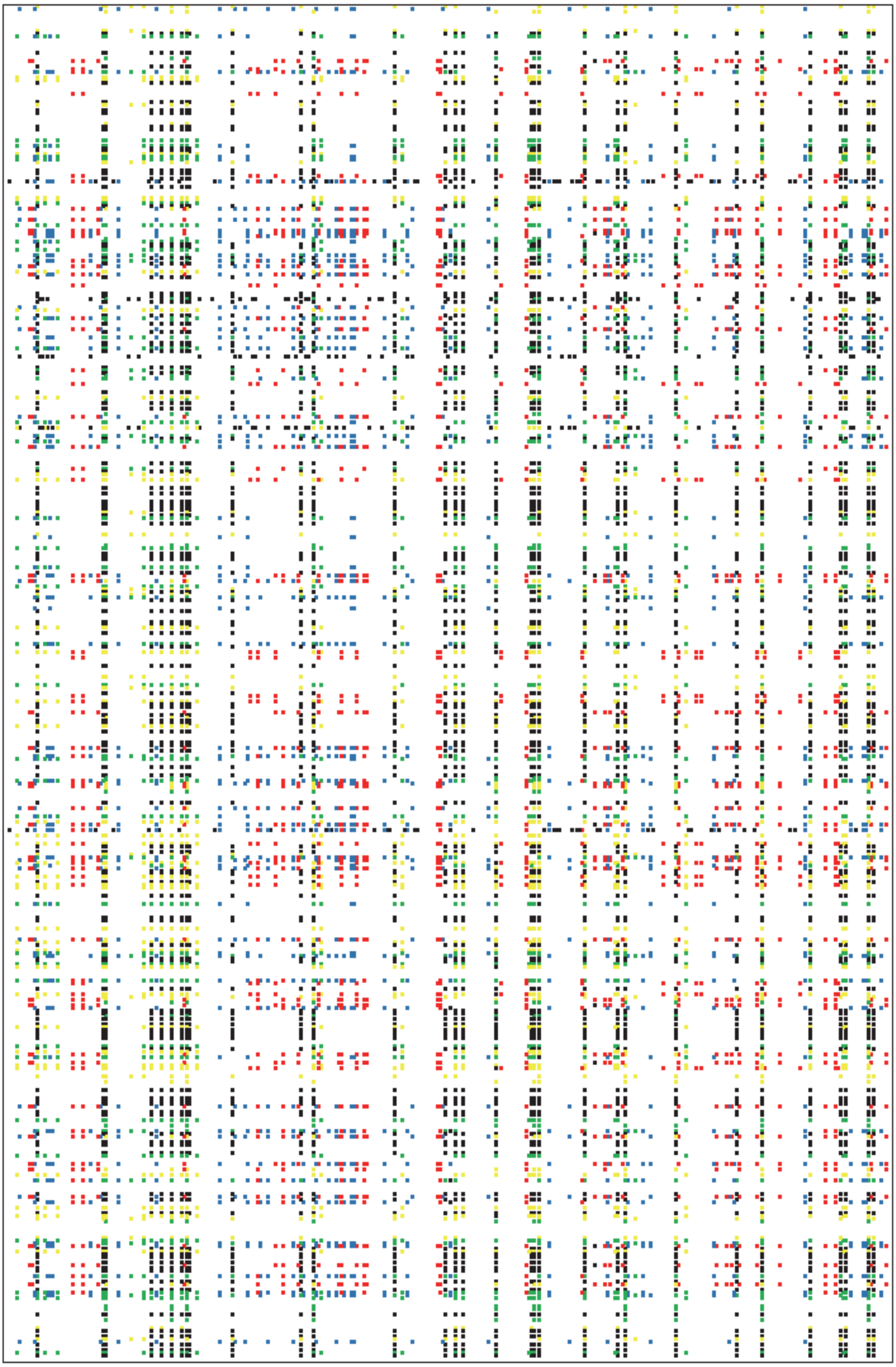


Figure 8 The coopetitive process tangled

The following algorithms are applied:

- *countones*—a fast O(*n* log *n*) heuristic for larger matrices, based on sorting by the frequency of ‘ones’;
- *conf*—conformity scale; *minus*—minus technique, and *plus*—plus technique—algorithms from the Monotone Systems metaheuristic (Mullat, 1976), (Vyhandu, 1980)
- *bea*—McCormick’s BEA, (McCormick et al., 1972)
- *roc2*—an enhanced rank order clustering (King et al., 1982)
- *modroc*—an extension of the rank order clustering for group technology (Chandrasekharan et al., 1986)
- *art*—a Carpenter–Grossberg neural network based clustering (Kaparthi et al., 1992) (Kusiak et al., 1991)
- *zodiac*—ideal seed method for part-family and machine-cell formation in group technology (Chandrasekharan et al., 1987)

As we can see, these originate from various domains such as different venues of operations research, group technology and cellular manufacturing. One often sought after goal is to perform block diagonal seriation (to permute the matrix at hand in such a way, that it would have a strong main diagonal), the meaning of which we will consider next (see Figure 9).

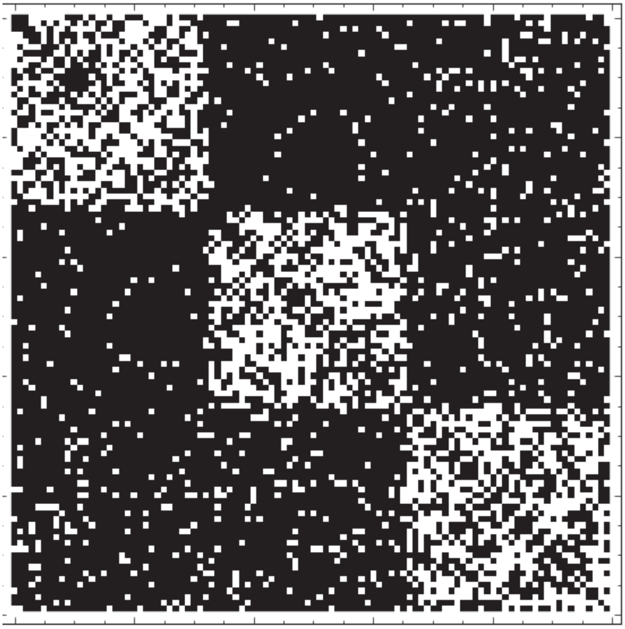


Figure 9 Matrix with a strong main diagonal

In cellular manufacturing (e.g. Burbidge, 1971), for example, the formation of groups on the main diagonal means that there are groups of manufacturing operations (machine-groups) that can physically be placed near each other in the product flow. The exceptions that lie off-diagonal, mark discrepancies in manufacturing, which disrupt the product flow.

In archeology, the co-appearance of the types of pottery (or the traits of pottery, e.g. the type of the handle) in graves was already studied more than a century ago (Petrie, 1899). Due to the fact that essentially only time-wise close objects coappear in a grave, a strong main diagonal was formed which marked the progression of time. The more distant the times to which the objects in a grave belong, the more off-diagonal the item would be, which is something that happens rather rarely. An entry in the table that is far off-diagonal would mark the seeming impossibility of an old pottery item having been placed in a contemporary grave.

In operations research (McCormick et al., 1972) calculated similarities between control variables in an airfield, and used the BEA algorithm (mentioned above) to permute the strong main diagonal, once again grouping the similar variables together and finding the groups.

In anthropology, skull group similarity was also already studied about a century ago (Czekanowski, 1909). As in the case of McCormick's work, the table at hand was essentially a symmetric similarity matrix. Both were two-way one mode matrices where, as stated above, we will only perform dependent permutations of rows and columns. The study of this small 13x13 table was performed by visual inspection and the result was once again essentially a block diagonal seriation, with the exceptional similarity traits lying off-diagonal.

As our student similarity matrix comes from SNA, we will once again see strong connections formed inside the groups lying on the main diagonal. The off-diagonal elements in social networks are ties that are formed between the communities, but in this case they are necessary. Thus, we not only have communities that are detected, but also relationships that span them – a fact which we are taking into account, when developing an indicator for coopetition.

Figure 10 puts forth an initial, unordered state of a small, two-community social network. The network has 36 nodes, and on the lower pane the final state of SNA is also depicted, while the two communities are permuted together on the main diagonal. As expected, there are a number of off-diagonal connections between the communities. We have seen that the situation concerning the interpretation of a similarity matrix of a social network has notable differences from that of the matrix which archeologists construct, for example. However, there are also similar traits. Thus, it is viable to cross-brush the algorithms of different domains.

2.2 Modularity maximization and permutations of Visual Matrix Explorer

The Visual Matrix Explorer tool lets us explore the adjacency matrices. As a result of our community detection, we will "organize" the matrix by constructing a block diagonal structure. We also need an objective function of our own in order to study the social network. In the previous section we chose to maximize the modularity of the social network.

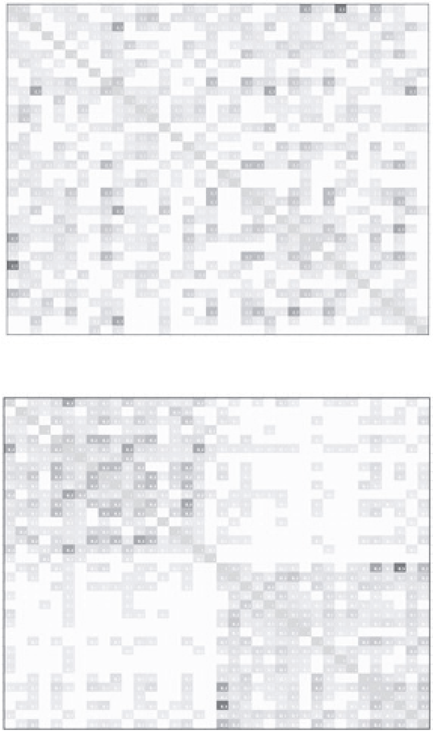


Figure 10 A 36 node social network adjacency matrix in unordered and ordered state, the emergence of strong block diagonal

Modularity therefore marks the communities that we permute together afterwards. This is an easy task once we know that a permutation **can** form a perfect block diagonal. In such a case the matrix preceding the permutation is in a “checkerboard” form. Figure 11 depicts the marked communities before and after being permuted to block diagonal form. It is only natural that Figure 11 has nothing off-diagonal, as the communities marked must form full cliques and can only be expanded as such.

In order to detect the communities of students that mark the departments in coopetition with each other, we already excluded some of the values in the adjacency matrix in section 1. The reasoning behind this was that although all values can be calculated in the adjacency matrix, whether they are still worth considering is another question, especially since our social network would therefore be fully connected – each student to every other one. It is definitely possible to surmise that some lower values of similarity might simply form noise that distracts us from our quest for retaining a signal in our data table.

Figure 12 puts forth our adjacency matrix from section 1 according to modularity maximization by deploying VME. The communities Marketing, Management, Accounting and Finance have been permuted together and selected in Red, Green, Blue and Purple colors, respectively. A bicolor version of this matrix would also underline the strong main diagonal that emerges, with less ties lying off-diagonal, connecting the communities. The students that are not included in any of the four communities have been grouped in the lower right corner.

Figure 13 shows another permutation of the same matrix, according to the Zodiac algorithm, as we apply a different objective function. The communities are depicted in an interleaved manner.

Next, we are going to drill down into the Red, Marketing community. Figure 14, on the left, shows the two subcommunities of Marketing, colored in lighter and darker red. On the right pane, we see the same communities re-permuted (with more degrees of freedom, as we have drilled down) according to the BEA algorithm (McCormick et al., 1972).

We see that the same two subcommunities are still perfectly separated. At the same time the majority of the off-diagonal lying connections between the two communities have been rearranged to the border of the communities. This way we now have both clustering into two communities, and an internal order within the two communities.

The ordinary algorithms for clustering have a shortcoming. They come up with clusters but the issue of the internal order of the leaves is not addressed – we have hereby overcome this (cf. (Liiv, 2010a)).

The BEA permutation is also closer to the block diagonal matrix form than the initial row order. At the borders of subcommunities thick chunks of data tying the two together can be identified. This has implications for our coopetition metric, which uses traditional conductance. In section 1 we described how traditional conductance takes into account the ratio of ties to the outside of the community to those to the inside. The internal order of leaves identified allows us to futher divide individual students. By crafting an indicator upon another indicator, we may divide community members into two – those who contribute more to the cooperative stance of the community (and have their personal traditional conductance above that of their community) and those who contribute more to the competitive stance of the community (and have their personal traditional conductance below that of their community). The BEA permutation brings out both, with students contributing to the competitive stance aligned to the borders of the two subcommunities.

Our next objective is to investigate the removal of some layers of the adjacency matrix and the amount of signal retained in the process.

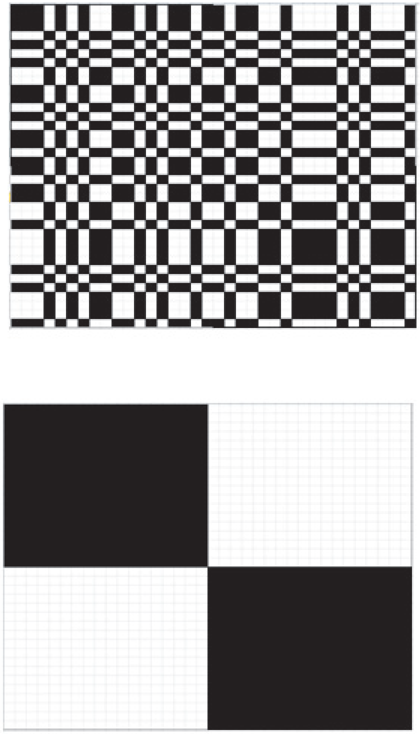
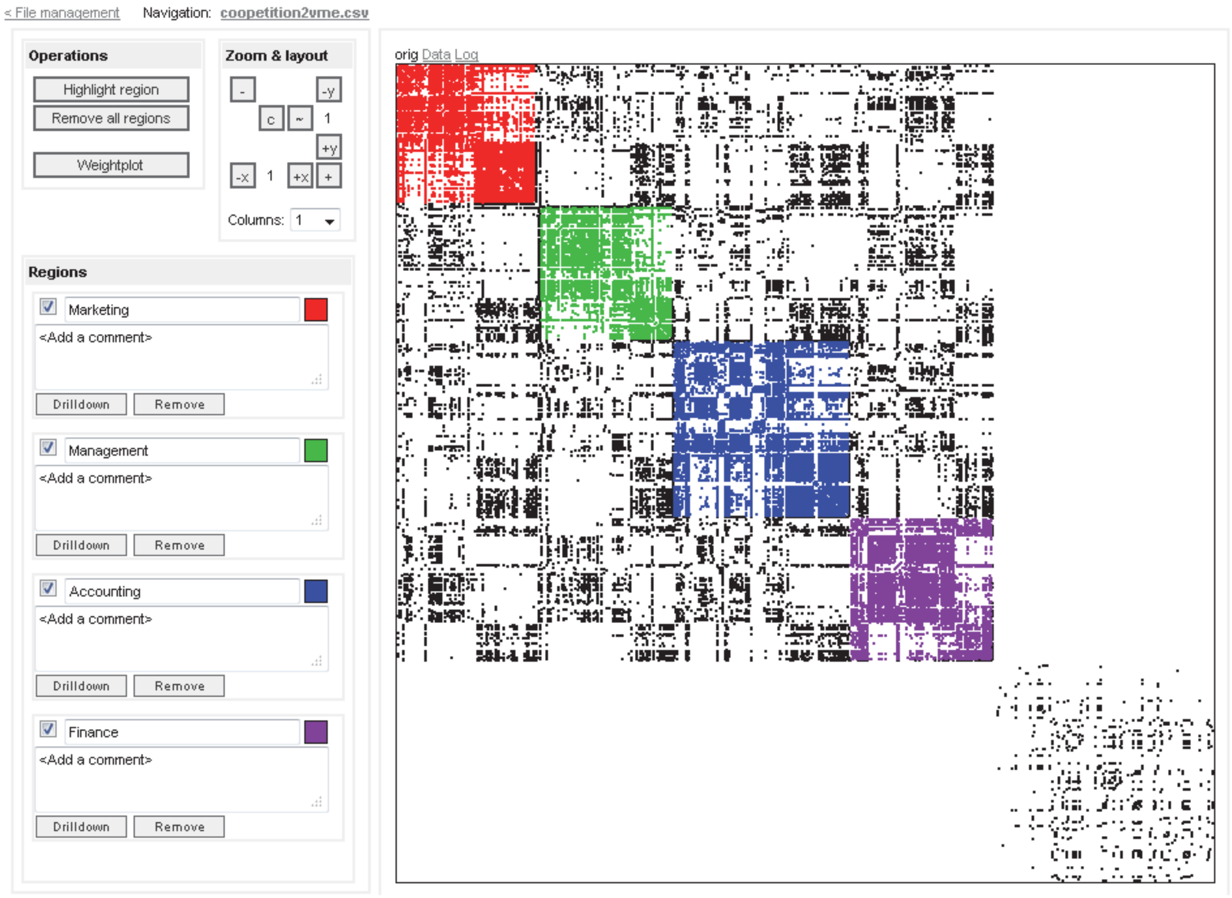


Figure 11 A matrix marking the communities of a 36 node social network – before (the "checkerboard" form) and after permutation



< File management Navigation: coopetition2vme.csv

Columns: 1

☑ Marketing 🟥

<Add a comment>

☑ Management 🟩

<Add a comment>

☑ Accounting 🟦

<Add a comment>

☑ Finance 🟪

<Add a comment>

Figure 12 VME user interface depcting the adjacency matrix of the social network of section 1

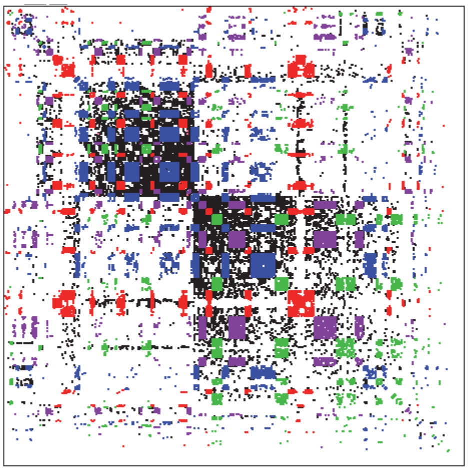


Figure 13 A permutation of the adjacency matrix depicting same communities in an interleaved manner

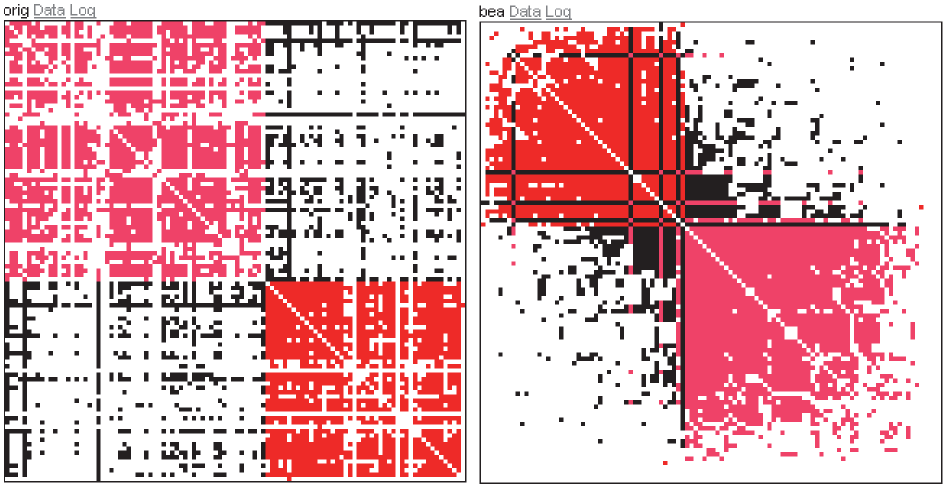


Figure 14 Drilldown of the Marketing community with fewer degrees of freedom – original essentially shows the modularity result, whereas BEA has rearranged the communities internally.

3 AN ANALYSIS OF COMMUNITY STRUCTURE DETECTION FOR EDUCATIONAL COOPETITION

The goal of this section is to study how the strictly optimal solutions of community detection, based on similarity matrices, depend on the parameter of the distance threshold setting method applied beforehand. In order to detect communities, we apply the often used modularity metric and arrive at strict optimality by linear programming. The distance threshold method is used, which makes the matrix more and more sparse. Thus, the best value of the threshold is determined by analyzing the number of subsequent clusters detected.

Our method is applied to educational coopetition data (from section 1) in the business school of TUT with four specializations, out of which we sample 36 students, selecting from a pair of specializations each time. Since the optimal number of clusters tends to be four, for any two-fold sampling we detect a natural division within each specialization as well (just as we did in section 1). The reason for this is a matter for further study. As a result, coopetition (the simultaneous competition and cooperation) between the departments of the business school is measured. Like in section 1, the average grade of the students is a proxy for the competitive score of the department. The traditional conductance is used as a proxy for the cooperative score of the department.

For our data, the optimal value for the threshold in community detection is 0.07. This way enough noise has been removed from the data but not too many values, so that vital information is retained. Thus, we most often obtain our goal of detecting four clusters in the two-fold sampling, which effectively displays the usefulness of fine-tuning the distance threshold while evaluating it through the strictly optimal community detection.

We also check for subcluster stability. Considering that our students of a given specialization have always been detected in two subclusters, has a specific student constantly ended up in the same subcluster? By finding the "checkerboard" pattern in the results we know the answer to be yes as this confirms subcluster stability.

3.1 Overview of the numerical method

Starting with early works such as (Zahn, 1971), network science has proposed a number of community detection measures and algorithms (Newman, 2004) (Chen et al., 2009). Modularity measure was put forth by (Newman et al., 2003) and it has been proven to be highly effective for community evaluation in practice (Danon et al., 2005). The strict optimization of modularity has been proven to be NP-hard (Brandes et al., 2008) and highly effective heuristics have been devised, e.g. (Blondel et al., 2008). This section aims to explore the relationship between strictly optimal solutions and parameters that are used for methods which refine the community detection data.

When visualizing the adjacency matrices of graphs, matrix reordering/seriation methods permute the rows of a community together, thus forming a strong block diagonal that visualizes the intracommunity ties with intercommunity ties displayed off-diagonal (see section 2 for a review and a software tool that implements a number of algorithms). A paper by (Wang et al., 2002) visualizes such a reordered shaded similarity matrix (by applying a method by (Gale et al., 1984)) and applies the distance threshold method with different parameter values by making the matrix more and more sparse (see Figure 15).

It is hoped that this way the irrelevant, immaterial noise is removed from the matrix, and only the information pertinent for community formation is retained. We aim to study such signal-to-noise ratio. The goal of the section is (by raising the value of the distance threshold parameter and at each step solving the NP-hard, strictly optimal modularity detection problem) to determine the optimal value for the threshold as we perform supervised learning.

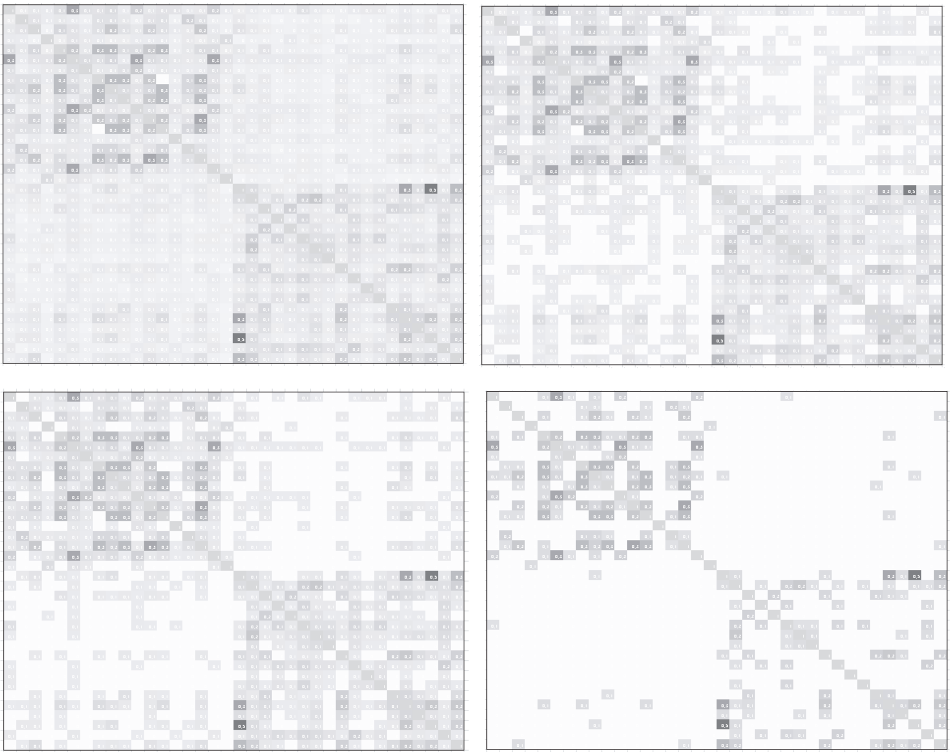


Figure 15 The use of the distance threshold method – a successively sparser adjacency matrix with its two communities permuted together. We study how this threshold influences the strictly optimal community detection.

3.2 Details of modularity and distance threshold method

(Newman et al., 2003), in essence, defines **modularity** as

$$Q = \sum(A_{ij} - E_{ij}) * X_{ij} \rightarrow max \qquad (3)$$

where A is the adjacency matrix of a graph, Eij is the expected value of the element of the adjacency matrix and Xij is a binary variablethat indicates whether the ith and the jth node of the graph are connected in the same community. The expected value, Eij, defined as

$$E_{ij} = \frac{\sum_{l=1}^{n} A_{il} * \sum_{k=1}^{n} A_{kj}}{\sum_{k=1,l=1}^{n} A_{kl}} \qquad (4)$$

is row sum*column sum/matrix sum. This means that modularity has essentially been carved after the Chi-Square characteristic, which also expects big values to appear in an element that has big row and column sums for example.

Communities are formed in such a way that, taking into account the whole matrix, nodes are connected where the respective value in the adjacency matrix is bigger than its expected value.

The work by (Brandes et al., 2008) on determining the NP-hard nature of modularity describes it as an integer linear programming (ILP) model. It has three types of constraints for the detected communities: reflexivity, symmetry, transitivity,

$$X_{ii} = 1 \qquad (5)$$

$$X_{ij} = X_{ji} \qquad (6)$$

$$\begin{cases} X_{ij} + X_{ju} - 2 * X_{iu} \leq 1 \\ X_{iu} + X_{uj} - 2 * X_{ij} \leq 1 \\ X_{ju} + X_{ui} - 2 * X_{ji} \leq 1 \end{cases} \qquad (7)$$

which ensure that only entire communities are selected. Specifically, the transitive nature of the problem ensures that if the first node is connected to the second, and that in turn to the third, the first also has to be connected to the third. Altogether, there are n^2 decision variables, n reflexivity constraints, n^2 symmetry constraints. The number of transitivity constraints is equal to the number of 3-combinations of n; but note that there are also three constraints that comprise each transitivity constraint. This way the total number of transitivity constraints is also equal to (the number of 3-permutations of n)/2 as the symmetric permutations can be pruned because of the symmetricity constraint. For example, a problem with 36 nodes will have 1296 binary decision variables and 22752 (36+1296+21420) constraints.

The **distance threshold method** has been applied in (Wang et al., 2002). This method essentially removes all distance values lower than a given threshold from the matrix. We gradually raise this threshold in subsequent iterations, and thus make the matrix more and more sparse, hoping to remove noise but not the signal from the data. In order to select the proper range for the distance threshold, the histogram of the data is constructed.

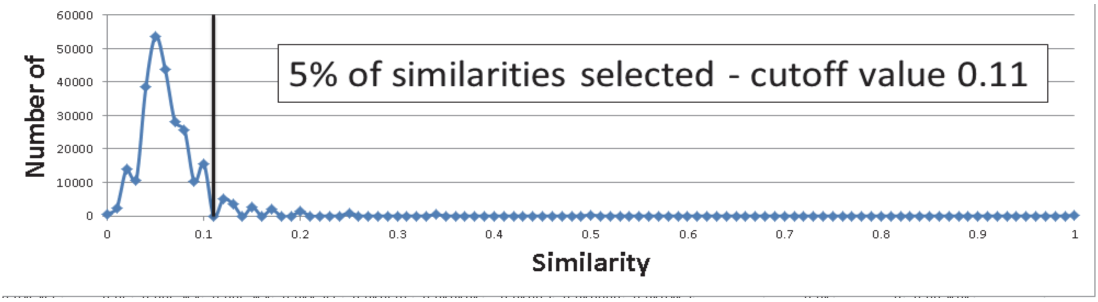


Figure 16 Histogram of the student similarity matrix – similarity being defined as the reciprocal value of the Hamming distance of the student-by-course matrix

3.3 Data on coopetition

The data on coopetition comes from the Tallinn School of Economics and Business Administration (TSEBA) of the Tallinn University of Technology (TUT). It consists of 331 students (a subset of data from section 1), who have graduated from the curriculum "Business" over the time span between 1997-2010. We have information regarding the courses that the students have attended, the grades they received, and the final departmental specialization chosen. Altogether, students have attended 759 courses. The course selection data follows the power law. The four specializations of TSEBA are finance, marketing, accounting and management (see Table 4).

Table 4 The students and the specializations of the business school TSEBA studied

Specialization	Number of students
Marketing	87
Finance	74
Management	71
Accounting	99
Total	**331**

As the first step of analysis a binary matrix (with rows representing the students and columns representing the courses attended) is constructed.

This matrix is about 95% sparse. Next, a 331x331 distance matrix between all the students is formed. At first, Hamming distances are used – counting the total number of different courses selected by the two students. The second step involves the calculation of the similarity matrix by taking the reciprocal value of the distance. In order to later select the threshold, we also calculate the histogram of the similarities (see Figure 16).

3.4 An application of the method

We begin by applying a fast modularity heuristic (Blondel et al., 2008) to the entire 331x331 similarity matrix in its original form (threshold 0), as well as with 95% of smallest values removed (threshold 0.11, cf. the histogram in Figure 16). While the full similarity matrix has badly detected the departmental specialization of students, resulting in 16 detected mixed communities; the sparse matrix ends up with 8 communities (see Figure 17). There, the students of each specialization are divided between two communities and all communities are comprised of only one type of students.

As heuristic modularity algorithms have potential community detection problems of intrinsic nature, we want to abstract ourselves from those and henceforth look at the detection of communities backed by the strict optimality of the modularity criteria.

We use Gurobi engine of Analytic Solver Platform and require it to solve well over thousand problems. Our chosen graph size has 36 nodes and thus, 1296 binary decision variables and 22752 (36+1296+21420) constraints. This problem has a solution time of roughly 90 seconds.

As there are only 36 nodes in the graph, we sample 18 each from two specializations for one optimization. There are six combinations for samplingtwo out of four specializations. We sample each specialization combination 8 times. We vary the threshold criteria from 0 to 0.12 by 0.01 steps. Thus at this step we perform 6*8*13=624 optimizations.

The results are depicted on Table 5, showing the average number of pure clusters obtained. In about 5% of optimizations the result had one cluster that was not “pure”, i.e. it contained more than one type of specializations. These optimizations are omitted from the results. For different threshold levels, the average number of clusters varies between 4.56 and 10.90, and is at its worst levels when the threshold is either too low or too high. Clearly the optimal threshold level for our data is 0.07.

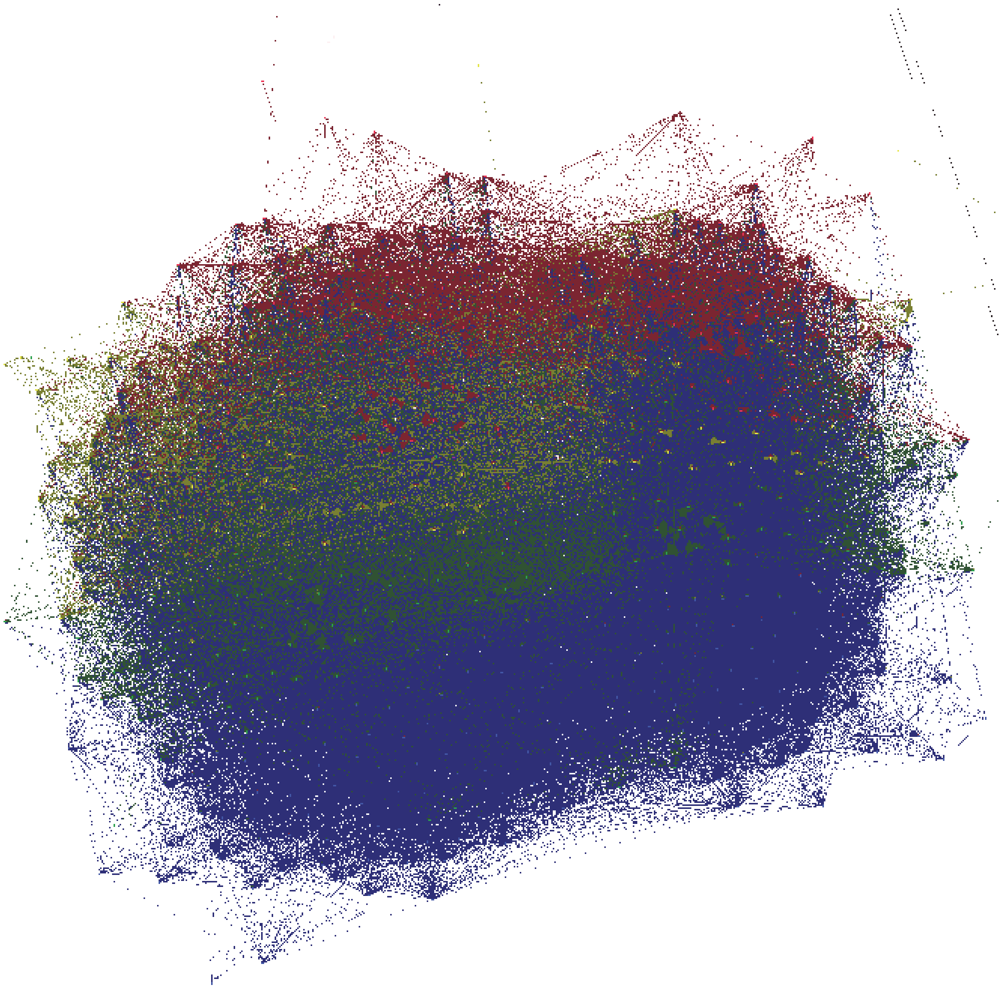


Figure 17 Coopetition between students, communities depicting departmental specializations and nodes marking students

Next, we look at the fact that there are four clusters in each optimization, not two. Simultaneously, the clusters are almost always “pure”. We study the subcluster stability within each specialization. In order to do that we set the threshold level at 0.07. We use all six combinations of the specializations. We sample each combination 100 times. This way we now perform 1*6*100=600 optimizations.

The results depicted on Figure 18 show very good stability. For all four specializations, the two internal subclusters (which can be seen from the “checkerboard” pattern) can easily be permuted together, leaving very little noise off-diagonal. The reason for the emergence of such internal subclusters, however, is a topic for a follow-up research. It can tentatively be surmised, though, that perhaps the reason is the revamp of the curriculum which was undertaken to move on with the Bologna Process.Coopetition score of the departments

Having detected the departmental specializations, we next measure their coopetitive stance, as wedid in section 1, by using the current data table.

Let us once more state that the competitive position of a department is evident in the average grade of its students. For the cooperative position we look at the traditional conductance of social network analysis (Mislove et al. 2007):

$$Cooperation = \frac{\sum linkages\ from\ the\ specialization\ leading\ to\ others}{\sum all\ intragroup\ linkages\ for\ the\ specialization} \quad (8)$$

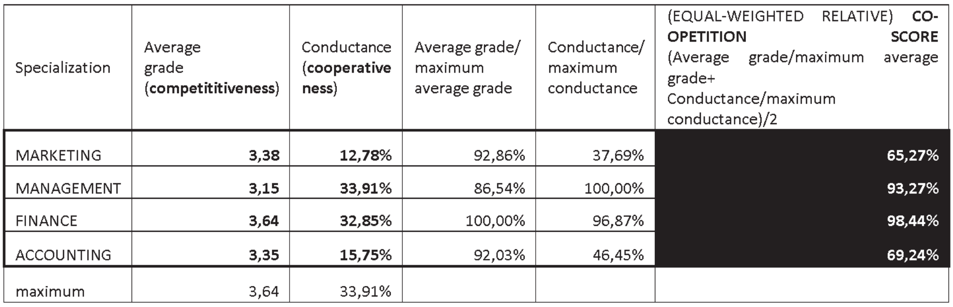
This uses students as proxies for showing how interlinked the curriculum is. If there are enough Management students taking Finance courses, it is going to be evident in the similarity matrix. We normalize both scores as percentages of the maximum, and calculate the coopetition score as an equally weighted average of the two.

The results are summarized in Table 6. The competitive position is the strongest for the Finance specialization. The curriculum of Management (which does worst in competitive terms), is the most interlinked with the others. The equal weighted score is the highest for Finance at 98%, but Management comes in as close second at 93%. Thus, our goal as a university would be to financially reward the coopetitive outcome, because we can ground our footing after detecting the communities in a sounder manner.

Table 5 The average number of „pure“ clusters obtained, dependent on the threshold selected, across all the two-fold sampling combinations of the four specializations; 18+18 students were selected for each optimization

Threshold	Specialization pair						
	Marketing-Finance	**Marketing-Management**	**Marketing-Accounting**	**Finance-Management**	**Finance-Accounting**	**Management-Accounting**	**Grand Total**
0	7,88	8,75	7,5	8,88	9,13	9,75	**8,65**
0,01	7,88	8,75	7,5	8,88	9,13	9,75	**8,65**
0,02	7,88	8,75	7,5	8,88	9,13	9,75	**8,65**
0,03	7,88	8,75	7,5	8,88	9,13	9,75	**8,65**
0,04	7,88	8,63	7,5	8,88	9	9,63	**8,58**
0,05	6	6,63	6	7	7,88	7,75	**6,88**
0,06	4,38	4,13	4,75	5,13	6	5,25	**4,94**
0,07	**4,13**	**4,25**	**4,5**	**4,63**	**5,13**	**4,75**	4,56
0,08	4,63	5,25	5	5,5	5,88	5,38	**5,27**
0,09	5,13	5,75	5,75	6,38	6,75	6,13	**5,98**
0,1	6,13	6,88	6,63	7,63	7,13	7,38	**6,96**
0,11	7,88	8,75	7,88	9,25	8,75	9,38	**8,65**
0,12	10,5	11,13	9,75	11,38	10,13	12,5	**10,9**

Table 6 The computation of the coopetition score



Specialization	Average grade (**competititiveness**)	Conductance (**cooperativeness**)	Average grade/ maximum average grade	Conductance/ maximum conductance	(EQUAL-WEIGHTED RELATIVE) **CO-OPETITION SCORE** (Average grade/maximum average grade+ Conductance/maximum conductance)/2
MARKETING	**3,38**	**12,78%**	92,86%	37,69%	**65,27%**
MANAGEMENT	**3,15**	**33,91%**	86,54%	100,00%	**93,27%**
FINANCE	**3,64**	**32,85%**	100,00%	96,87%	**98,44%**
ACCOUNTING	**3,35**	**15,75%**	92,03%	46,45%	**69,24%**
maximum	3,64	33,91%			

3.5 Concluding discussion

This section has combined the strict optimization of the modularity metric with the use of the distance threshold method which shows the importance of discarding the noise but retaining the signal. As the problem is NP-hard, data has been repeatedly sampled from the whole set. The communities detected this way have proven to be highly stable and more than 99% of all the detected clusters have been "pure", i.e. composed of only one specialization. We found that the optimal threshold level for our data is 0.07. Meanwhile, the detection of clusters was performed in a fractal-like self-similar fashion, discovering them on a small scale and finding validation in the whole data set. The detection of communities has served the purpose of calculating the coopetition score for a department. This proved to be dependent upon competitive well-being, like in the case of Finance, but also strongly influenced by how well the department serves the rest of the business school, like in the case of Management.

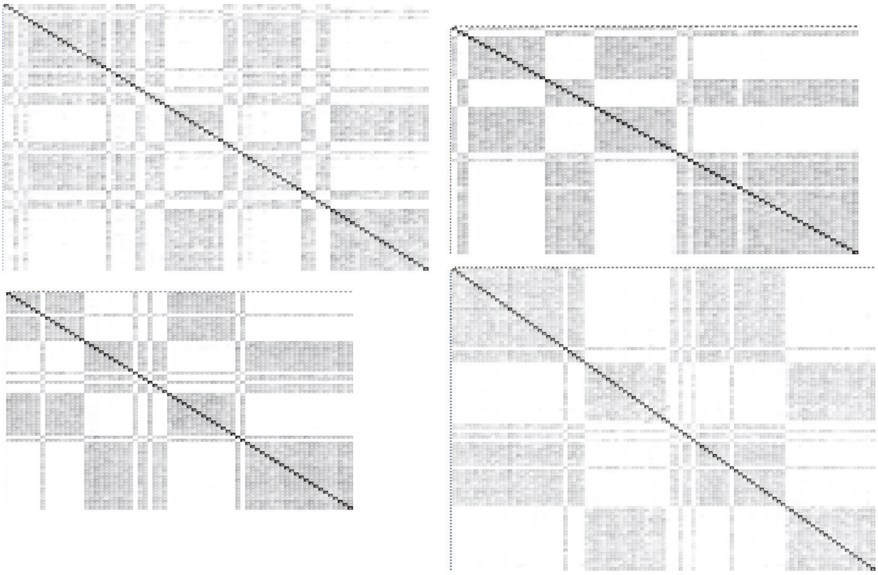


Figure 18 The subcluster stability for the four specializations – Marketing, Finance, Management and Accounting, each matrix showing the “checkerboard pattern”, which can be permuted together, containing two clusters and very little off-diagonal noise

4 STUDENT CHURN IN THE LIGHT OF THEORIES ON BUSINESS RELATIONSHIPS

This section presents a discussion on the applicability of business theories to student retention research. After giving a motivation for and a review of previous retention research, relationships in business networks are compared to those between a student and a university, putting forward relevant characteristics in the context of TUT case study. Theories regarding the taxonomy of customer churn, its determinants and consequences are also viewed in this context. The implications for educational data mining, for both strands of research, are forward-looking, as we are setting the grounds for quantitative analysis of student retention in the subsequent section.

4.1 Introductory discussion

Student retention rate is one of the key indicators of success that universities have as regards the graduation of its students. A high retention rate is important for the stability of institutional budgets. It is also a tool for measuring national institutional accountability. Internationally, the ranking systems of universities that measure the quality of the provided education take the retention rate into account. The rate of graduation of students has become especially important over the last half century, as the number of people attending universities has risen in such a way that 90% of the scientists of mankind were alive in 2002 (Economist, 2002). In developed countries, student churn is as low as 8% in the top universities, about 35% in the average case and as high as 50% in the case of "open enrollment" (Devarics and Roach, 2000).

The goal of university intervention programs and services (UIPS) is to help students develop the necessary skills to graduate. Intently bringing a parallel from economics, one of the goals of a university could be stated to be the maximization of the number of credits "sold" to its customers, with a notion that it is the students who receive the credits and who thus have to be germanely applicable for them. The decision by UIPS to intervene is in turn made in an economic context. First, the students potentially dropping out have to be identified. If a student who has high potential for graduating was to churn, nearly 180 credits would be lost to the university. It is cost effective to concentrate the effort first on high potential students because, on the average, they can also be helped with the least amount of resources.

However, as far as the intervention process in its entirety is concerned, it has so far seemed almost impracticable to justify the effort in cost benefit terms. Namely, the indicators of success – churn rates – have not changed much as a result of UIPS (Seidman, 2005).

Therefore an imperative of this work is to identify students in need of help more successfully. This survey will aim to identify additional uses for secondary information as regards students that can be put to use in educational data mining, all in order to better position the university in this customer-driven form of accountability.

Subsection 4.2 will explore the motivation for considering new perspectives of the student-university relationship. Subsection 4.3 will give a brief overview of the research on student retention thus far. Subsection 4.4 will elaborate on the student-university relationship in the light of theories on business networks with the aim of providing insights into educational data mining. Lastly, subsection 4.5 will do the same for theories regarding customer churn, followed by a concluding discussion in subsection 4.6.

4.2 A motivation for considering new perspectives of the student-university relationship

The rationale behind this work is the need for universities to manage its relations with students more actively. Amongst the different relationships managed in a society, the ones relating to a university take place in perhaps the highest venue of intellectual capital. While there are differences when compared to the management of the relationship in the business domain, there are certainly also similar traits.

The initiation of the relationship and avoiding churn in TUT. The relationship starts with universities recruiting high school students. For TUT in its most intensive cases, this takes the form of partnership agreements with leading schools, some of whom have a university program integrated into their curriculums. Examples of other initiatives during the "selling effort" are Technology School (TUTTechSchool, 2014), the attendance of which starts in junior high, and City Camp. Guest lecturers visit high schools, there is also a Summer School and visits to TUT are organized. The newly founded Modern Estonian Knowledge Transfer Organization for You (MEKTORY) (TUTMEKTORY, 2014) (that handles the contacts with the industry) also regularly hosts guests form schools. Furthermore, TUT takes part in the development of high school curriculum, e.g. it recently finished a three module block of courses on informatics to be taught throughout the high school years (Mironova et al., 2013). As a result of those activities, TUT has an ongoing relationship with teachers. Student status is attained by placement as a result of national high school graduation exams and the university offers preparation courses for those.

In managing student churn, the biggest danger looms in the first year of attendance (Schertzer et al., 2004). Therefore, the academic year starts with a preparatory week for freshmen. Also, older students can become tutors who provide additional help to newcomers during the first part of the semester. There is an e-course "Self-Management" offered to freshmen as well. In this tech school, there are a few key courses (mathematical analysis and physics) that seem to be the best predictors of future success or failure from anecdotal evidence, but this is something we also seek to quantify by applying feature salience detection in the next section. Preparatory courses are offered for those subjects. Recently, the position of "group counselor" was created in some faculties. This is an additional assignment (besides teaching) for a member of the faculty, something that has been revived from the past and in a way resembles a high school teacher's assignment. Other faculties are using mentors, which is essentially a different name for the same phenomenon. Also, attendance has been made mandatory in the first semester during some years, and information on midterm results is gathered. As a result, faculty administrators contact students who are falling behind by phone. Another new trend is the postponement of some general courses until second year, thereby opening the opportunity to study specialized (less fundamental) courses sooner.

Having contemplated the initial contact between the university and the student from the university's point of view, we are next going to assume that of the student. Theories in economics (Caves and Porter, 1977) mention barriers to entering into market relationships and, as we will see, entering into studenthood comprises of overcoming certain entry barriers for a freshman as well.

On the one hand, studying at a university is qualitatively harder than it was in high school. Students often feel that the university curricula are substantially different from that of a high school. The amount of independent work is greater and the pace in classes is faster.Furthermore, many of the general courses are considered to be amongst the most difficult at the university. It takes time to get accustomed to the administrative system, procedures and to university life in general.

These **entry barriers** are complemented by the changes in personal life. Former high school students now have complete responsibility for planning their schedules. Meanwhile, most have started living on their own (rather limited) budget. On the average, across the entire study period, 60% of students are working concurrently with the studies (Teichmann, 2010). On the other hand, only 10% of students have a monthly budget exceeding $600 (Teichmann, 2010). Freshmen, now at the helm of their lives, have increased responsibility for coping with stressful situations. They have to balance their emotions and continue learning to resolve conflict situations (Coelho et al., 1963). One of the main skills acquired at university is the ability to learn. Students acknowledge the need for developing self-discipline, self-motivation, ability to learn for examinations; uncertainty about one's ability to cope and uncertainty regarding making the correct career choice is often present.

Together with entry barriers and changes in personal life, we see that as the student invests not only his energy, but also money into that relationship, these create **sunk costs** and a strong incentive for staying at university.

As evidenced from this discussion, we have been able to draw parallels with the business domain as regards both initiating contact with the student and the process of avoiding churn, which provides the motivation for this study. We need to understand the determinants of this relationship more thoroughly, both from the viewpoint of the university and the student, and draw additional parallels between this person-to-institution relationship and business-to-business relationships that occur on the marketplace.

While earlier work on that topic has focused on finding parallels between customer relationship management and student relationship management (Ackerman and Schibrowsky, 2008), our focus is different. We juxtapose customer retention and churn theories with the state of affairs at universities. Our second goal is to use the theories regarding business networks to find parallels with the relationship between universities and students.

4.3 Theoretical foundations of student retention research

According to (Seidman, 2005), previous research has traditionally focused on describing the economic, organizational behavior related to, and the psychological and sociological factors pertaining to student retention.

Economic factors have customarily been considered in the research on student retention. According to human capital theory, the choice to study is made in terms of cost-benefit analysis. If all in all, the amount of income with education is greater than without it, then education is pursued (Tinto, 1986) (Braxton, 2003).

Organizational behavior issues. (Bean, 1980) has drawn parallels from the theories of employee turnover. The factors influencing student retention are therefore routinization; participation; instrumental communication; integration; distributive justice; grades, practical value and development which represent the pay in these studies; courses which represent the job content; membership of campus organizations as professionalism; kinship, marriage; and transfer to another university as an opportunity.

This study is very close to ours, as it looks at the employee turnover juxtaposed with the choice that a students faces, when deciding whether to withdraw from the relationship. It finds similarities, just like we do, with business network relationships and customer churn factors.

Psychological factors. (Bean and Eaton, 2000) have considered student's past behavior, beliefs, normative beliefs and how these influence the way the university is perceived. The positive development scenario in that case includes development of psychological characteristics like positive self efficacy, declining stress, increasing efficacy, internal locus of control. Students interact with the university as well as the external environment, e.g. parents and spouses. The psychological processes lead to academic and social integration, institutional fit and loyalty, intent to persist and persistence. (Astin, 1997) finds that involvement is the key to retention and defines it as the amount of physical and psychological energy devoted to the academic experience.

Sociological factors. (Tinto, 1986) paper on students' interactions with the environment finds that student peers, family socioeconomic status, mechanisms of anticipatory socialization and the support of spouses constitute factors which influence the departure decisions of university students'. (Bean and Metzner, 1985) state that in the case of nontraditional students, e.g. older students, environmental factors (such as hours of employment, family, finances) have greater impact than academic variables, such as course availability. They also state that most important retention variables likely differ for groups such as older students, part-time students, ethnic minorities, women, academically underprepared. (Berger, 2000) discusses cultural capital in the form of informal interpersonal skills, manners, linguistic and educational credentials and investigates the possible mismatch between that of a student and the university.

(Kuh and Love, 2000) investigate cultural traits, e.g. the importance attached to attending university, cultural distance from the university and the importance of achievements and persistence as cultural traits in achieving the persistence of a student.

4.4 Implications of research on business networks

First, we focus on the close-knit relationships in business networks. We are accustomed to looking at relationships between entities as either following a **strict hierarchical fiat** (in the case of the parties belonging to the same organization), or essentially being **arm's length transactions** (in the case of a market relationship between the entities). However, recent organizational theoretical conceptualizations as well as the works on relationships between companies describe a more nuanced reality. Drawing parallels with the most complex examples in nature, e.g. the human brain, and wider changes in many fields of research, (Hedlund, 1986) finds that today's corporations are heterarchical. (Burgelman, 1983), (Burgelman, 1994) discusses autonomous strategic initiatives that take place in corporations, which violate the principles of hierarchy.

On the other hand, the relationships between companies are notof a pure arm's length transaction type either. Works of (Forsgren and Pedersen, 1998), (Forsgren and Pahlberg, 1992), (Forsgren and Johansson, 1992) and (Forsgren, Ulm and Johanson, 2005) show that not only is a corporation essentially a network of units (as is elsewhere put forth by (Ghoshal and Bartlett, 1990), cf., from a multinational corporation subsidiary perspective, (Birkinshaw and Hood, 1998)), but it is also embedded in a business network of its own.

Next, we are going to concentrate on the business network perspective propounded by the school of Uppsala, but before that one further reference is warranted. As our objective is to view the relationship of a student with the university in the light of these conceptualizations, it is important to state that the relationships between companies, as well as within them, have an even wider number of manifestations. As was more thoroughly elaborated in section 1, big multinational corporations have also been stated to be comprised of internal markets, with an ongoing internal competition for world product mandates, centers-of-excellence and the likes between the subsidiary units (Birkinshaw, 2000), (D'Cruz, 1986), (Andersson and Holmström, 2000); and the supply chain relationships that a firm belongs to have been described as coevolving systems (Varga, 2009) (See also (Eisenhardt and Galunic, 2000)).

Again, our goal is to view these business-to-business and intraorganizational relationships, with a goal of finding similarities when it comes to the relationship of a person with an organization. Considering that this work (in section 2) quantifies the coopetition that occurs between university departments for students, we also definitely look for coopetition as something that can occur between students throughout their university studies. This issue is further addressed in the discussion about future research.

Returning to the business network perspective, according to the body of work from Uppsala University, two firms gradually increase their **commitment**, a key characteristic of business relationships, as they do business with each other. A process of **learning** about each other's capabilities, needs and strategies occurs, as well as a **formation of routines** for undertaking transactions. As sides **adapt** to each other incrementally, business networks usually encompass mutual adaption, which can only sometimes be unilateral.

Knowledge transfer is also a key characteristic of business networks, as organizational learning taking place. The result of knowledge development that occurs in business networks is often **tacit and intangible**. This means that subsequent changes (e.g. depletion) in the capabilities of an organization go largely unnoticed to an outside observer at first.

A juxtaposition of this theory with the state of affairs in university settings creates many parallels. In its essence, the relationship between a student and the university is also somewhere between hierarchical fiat (HF) and an arm's length transaction (ALT). The option of simply buying a right to study in single classes from the Open University of TUT would be most similar to ALT. For a university, the studenthood has a certain fixed length of nominal years to be concluded with the attainment of a prominent status of being able to call the institution one's *alma mater.* One of the goals of this status, for a university, is populating the ranks of its faculties. This can chiefly reach the HF in the case of a person becoming one of the faculty members.

Most of the relationships, however, are somewhere in-between. For example, it is common for students to pay for part of their studies by giving consults to their peers, all occuring within the administrative framework of the university. Becoming a teaching assistant and simultaneously teaching in the latter stages of one's studies is an even more common course of action.

The steep entry barriers and the commitment it takes to enter the university relationship that we considered above are other traits in common with relationships in business networks. As stated above, this forms a sunk cost that makes it unprofitable to exit the relationship prematurely. Being a university graduate often enables one to pursue the career goals that have been previously undertaken, which would have been denied in the case of churn. We can also say that the entry barrier and sunk costs explain the fact that student churn lessens considerably after a successful entry into studenthood.

As far as adaption is concerned, ideally it is mutual. Students, who are usually at least a generation younger than the tenure, are an active factor in introducing the applications of new technologies (e.g. videocasts, the use of social network communities etc.) to the classroom. This results in knowledge transfer and organizational learning on the behalf of the university. Moreover, student feedback in assessing the courses is a driver for the university in keeping itself abreast of times, and the university adapts its formal procedures as a result of student feedback.

On the other hand, the students also adapt. Due to the fact that the adaptation to the procedures of the university is incremental, it is natural that handling the relationship with the student in the latter stages takes fewer resources. More importantly, it can be stated that the student adapts their profile to that of the university. The student chooses research papers and theses topics that are aligned with the strengths of the university. The student also develops working relationships with those members of the faculty who shape the profile of the university.

This touches upon the much emphasized topic of integration the university with the industry. As the relationship between a student and the university comes to a successful conclusion, these personal ties are carried on as contacts with the industry, thus forming a basis for company-university relationships. This advocates for the argument about the importance of permanent faculty because once again intangible assets are formed and should not be depleted.

4.5 Implications of research on customer retention and churn

Our second focus is on the theories of customer retention and churn. Customer churn is defined as the propensity of customers to cease doing business with a company in a given period of time (Qian et al., 2006). In the current work, the term "student churn" is used interchangeably with the terms "dropout rate" and "attrition rate".

It is often stated that the cost of acquiring new customers is several times higher than that of retaining existing ones (e.g. Yankee group, 2001). Since the main goal of business entities is the maximization of shareholder value, linking customer retention to net income (the bottom line) is also relevant. This is done by measuring customer lifetime value (CLV), which is essentially the time adjusted monetary value of all the dealings with the customer. The CLV has also been found to be progressively related to customer tenure (Reichheld and Sasser, 1990), thus furthermore underlining the importance of customer retention.

Hence, in the work regarding student relationship management (akin to customer relationship management) by (Ackerman and Schibrowsky, 2008), the CLV of students is calculated. The authors make a basic mistake in calculating the present value of revenue rather than the net income. However, the calculations do underline the importance of being able to retain students all the way through graduation.

According to previous works (Villanueva and Hanssens, 2006), the most important determinants of customer retention can be classified as **switching costs, customer satisfaction and future considerations**. According to (Klemperer, 1987) there are at least three types of switching costs, the first being **transaction costs**. As regards the relationships in business networks, we did not consider the scenario of changing business partners, which is intrinsically more relevant to the discussion on churn. If a student decides to change the university he attends, transaction costs related to the procedure occur. While the percentage of university graduates who have attended more than one higher institution has been estimated to be above 35% in the U.S. (Seidman, 2005), Estonia only has a small number of relevant universities, and therefore the process seldom takes place.

The second type of switching costs is the **learning costs**. One will undergo these when choosing another university as the curricula are almost never fully compatible. The third type is the **artificial/contractual costs**, which give a student an artificial incentive to stay with the university, e.g. scholarships.

As far as satisfaction is concerned, studies linking it to retention have, in general, found a link between the two (Villanueva and Hanssens, 2006). In university settings, it is common practice to gather the assessments of students of the courses they took. In TTU's case, while on a certain level the process has taken place for years, the introduction of Study Information System (OIS) now enables mandatory student assessments. We already mentioned the future usage and expectations regarding the successful completion of the studies in the context of business networks.

According to a taxonomy proposed by (Mattison, 2005), churn can either be **internal or external**. External churn can in turn be **involuntary or voluntary**. Involuntary external churn happens in business when a contract is cancelled for reasons such as death, fraud, bad debt or underutilization. In the case of a university student, this would take the form of exmatriculation, if the pace of studies is too slow, or due to accruement of education funding debt. The reasons for voluntary external churn are alike in business and university settings – it takes place because of relocation, a switch to a competing university, a change to an alternative career path due to family considerations, etc.

External voluntary churn is additionally classified as either **deliberate or incidental** – deliberate reasons are grounded in dissatisfaction with the service and incidental ones with the changes that take place due to considerations regarding additional, fundamental decisions about one's future.

Internal churn deals with changes that take place within the university framework, e.g. an external or part time student becoming a full time student or vice versa, when insufficient performance results in a part time study plan. The types of churn mentioned above can additionally be classified as either **customer or competitor initiated**, with the majority of changes belonging to the former group in the university context.

4.6 Concluding discussion

When discussing churn, we found parallels between the taxonomy induced from business settings and the state of affairs at universities. We also found that relations that are developed in organizational networks closely resemble the student-university relationship as far as the strict hierarhical hiat/arm's length transaction principle, entry barriers, sunk costs, commitment, learning, formation of routines, mutual adaptation, knowledge transfer, and tacit and intangible assets are concerned.

In addition, the determinants of retention, such as switching costs (transaction costs, learning costs, artificial/contractual costs), satisfaction and future considerations were similar for the retention of students and customers alike.

Our next goal is to apply data mining in order to quantify student retention. A further point to be made would be about the need to differentiate the students in the course of data mining, as it was done in CLV analysis regarding churn. Specifically, there is a need to pinpoint the student retention effort because UIPS will need to allocate its resources with maximum effect. This is exactly what was done in TTU in 1970s, when a data mining project was undertaken. The study of (Lõhmus, 1974) used a method of contrasts in order to identify churning students who were to be helped ahead of others.

5 PREDICTING STUDENT RETENTION BY LINEAR PROGRAMMING DISCRIMINANT ANALYSIS

The goal of the section is to predict student retention with an ensemble method by combining linear programming (LP) discriminant analysis approaches together with bootstrapping and feature salience detection. In order to perform discriminant analysis, we linearize a fractional programming method by using Charnes-Cooper transformation (CCT) and apply linear programming, and compare this with an approach that uses deviation variables (DV) to tackle a similar multiple criteria optimization problem. We train a discriminatory hyperplane family and make the decision based on the average of the created histograms, thereby reducing the variability of predictions. Feature salience detection is performed by using the peeling method, which makes the selection based on the proportion of variation explained in the correlation matrix. While the CCT method is superior in detecting true-positives, the DV method excels in finding true-negatives. We obtain optimal results by selecting either all 14 (CCT) or the 8 (DV) most important student study related and demographic dimensions. We also create an ensemble. A quantitative course along with the age of the student at accession are deemed to be the most important, whereas the two courses resulting in less than 2% of failures are amongst the least important according to peeling. A five-fold Kolmogorov-Smirnov test is undertaken in order to help university staff in devising intervention measures.

5.1 Overview of the method

The classification methods used in data mining include discriminant analysis, support vector machines, decision trees, neural networks, gene expression programming and others. Linear or non-linear techniques are applied; methods such as self-organizing maps, artificial neural networks either alone or in ensemble are combined with graphs, etc. As linear methods generally perform faster (Bajgier et al., 1982), we seek to apply the simplex method of linear programming to perform the linear discriminant analysis. We use this instead of the oft-chosen Fisher's linear discriminant.

A number of approaches have been proposed for this, including (Bal et al., 2006), (Erenguc et al., 1990), (Freed et al., 1981), (Glover 1990), (Koehler 1990), (Kou, et al., 2003), (Retzlaff-Roberts, 1996). One of our contributions is following a call (by (Retzlaff-Roberts, 1996)) for applying variations of Data Envelopment Analysis (DEA), and deploying Charnes-Cooper transformation, the use of which is central in DEA. We combine this with another, more advanced LP discriminant analysis approach that deploys deviation variables (Kou et al., 2003).

We perform feature salience detection by applying a method called "peeling". Methods such as principal component analysis (PCA) identify new orthogonal dimensions in data and reduce dimensions. We seek to only do the latter by selecting the "more important" variables in the original data – those that account for the major proportion of the variation.

In order to predict student retention, we apply data mining.. We preprocess the data and use it for training our model to determine the result by using only 7 to 14 variables and 300 cases. This allows us to apply bootstrapping and create a family of discriminatory hyperplanes. Each case will then correspondingly have a histogram of its estimations, on the average of which we will base our decision. It will be shown that such averaging will enable us to take better decisions, as results vary significantly from one bootstrapping iteration to another.

As a subsequent step of applying feature salience detection, bootstrapping and two methods of LP discriminant analysis, we also discuss the merits of using an ensemble approach and combining the deviation variable and the Charnes-Cooper transformation classification methods.

Finally, we apply a five-fold Kolmogorov-Smirnov test on the histograms to find out how certain we are about students' future. Our recommendation will serve the staff members responsible for taking future action.

5.2 Multiple criteria linear programming (MCLP) approach for two class discriminant analysis

5.2.1 Problem description

Linear discriminant analysis is used in order to find a hyperplane that separates the two sets of students in the best way achievable. Thus we have:

$$\begin{aligned} A_iX \le b_i, A_i \in G_1 \\ A_iX \ge b_i, A_i \in G_2 \end{aligned} \qquad (9)$$

In order to do that, we find the correct and the erroneous distance that each data point has to the hyperplane (see Figure 19). Correct distances are denoted by β-s and erroneous distances by α-s, which we add to each equation. The signs of β and α will depend on whether the student will actually drop out or graduate. Thus our objective is to simultaneously maximize the sum of β-s and to minimize the sum of α-s. We arrive at the following set of equations (note that for one student, only either α or β will be different from zero, depending on the accuracy of the prediction, or both will be zero if the student lies on the discriminatory hyperplane):

$$\begin{gathered} Z_1 = \sum \beta_i \rightarrow \max \qquad (10) \\ Z_2 = \sum \alpha_i \rightarrow min \\ A_iX = b_i + \alpha_i - \beta_i, \forall A_i \in G_1 \\ A_iX = b_i - \alpha_i + \beta_i, \forall A_i \in G_2 \\ \alpha_i, \beta_i \ge 0 \end{gathered}$$

Note that some training set cases will indeed be placed on the hyperplane due to the nature of the linear programming methods used. However, the results still only have a binary interpretationas the validation and testing cases virtually never lie on the hyperplane.

5.2.2 Charnes-Cooper transformation approach to the MCLP

In order to solve the problem with the simplex method, we need to have a single objective function. We use the following fractional programming problem, where the numerator is essentially a vector consisting of all β-s and α-s, and the coefficients of α-s are zero; and the denominator is a vector consisting of all β-s and α-s, and the coefficients of β-s are zero. Once again we add β-s and α-s to inequalities of the constraints with corresponding sign changes (not explicitly indicated in the following equations).

$$z = \frac{cx + c_0}{dx + d_0} \rightarrow max \qquad (11)$$
$$Ax \leq b$$
$$x \geq 0$$

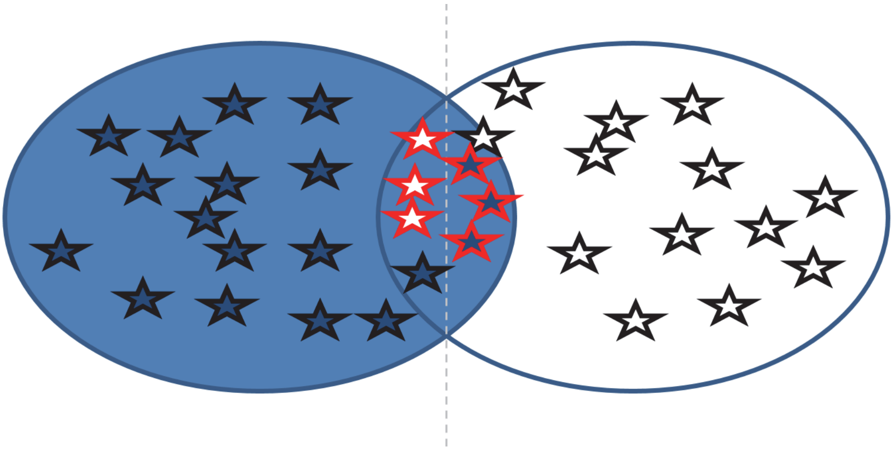


Figure 19 Separating hyperplane that determines the correct and erroneous distances of data points. The color of the boundary line indicates whether the person has been successfully identified as either graduating or dropping out (black – successful identification; red – incorrect assessment) and the color inside the star indicates whether the person has really graduated or dropped out (blue – graduated; white – dropped out).

The LDA models close to this were discussed in (Retzlaff-Roberts, 1996) and at the end of the article an idea of using Data Envelopment Analysis was put forth. We take up this call and linearize the problem by applying Charnes-Cooper transformation, the use of which originated from DEA in the following way. We denoto y_0 as:

$$y_0 = \frac{1}{dx + d_0} \qquad (12)$$

and x_j as:

$$x_j = \frac{y_j}{y_0} \qquad (13)$$

Thereby arriving at the following single criteria linear planning problem:

$$z = cy + c_0 y_0 \rightarrow max \qquad (14)$$

$$\begin{gathered} Ay \leq by_0 \\ dy + d_0 y_0 = 1 \\ y \geq 0 \end{gathered}$$

This has decision variables that have been transformed and an additional constraint that equals 1. Traditionally we would need a constant d_0 that avoids the possibility of dividing by zero but not in this case because there is virtually no chance of our data being discriminated perfectly. Further on we will use the variable y_0 for configuring the Charnes-Cooper transformation LDA model.

5.2.3 A comparative approach to the MCLP discrimination – the use of deviation variables

Paper (Kou et al., 2003), proposes another, more advanced approach to data mining MCLP discriminant analysis problems. It uses deviation variables, which take the absolute value of the left side of the equation. As sums of α-s are added to, and sums of β-s are subtracted from a constant, we can now include both in a single objective function, as follows:

$$\begin{gathered} Z = d_{\propto}^{-} + d_{\propto}^{+} + d_{\beta}^{-} + d_{\beta}^{-} \rightarrow min \qquad (15) \\ \propto_* + \sum \propto_i = d_{\propto}^{-} + d_{\propto}^{+} \\ \beta_* - \sum \beta_i = d_{\beta}^{-} + d_{\beta}^{+} \\ A_i X = b_i + \alpha_i - \beta_i, \forall A_i \in G_1 \\ A_i X = b_i - \alpha_i + \beta_i, \forall A_i \in G_2 \\ \alpha_i, \beta_i \geq 0 \end{gathered}$$

The parameters to be configured in this model are the constants β* and α*.

5.3 Feature salience detection with peeling

Methods such as the principal component analysis (PCA) reduce dimensions by retaining most of the variability in the data. The peeling method, (Vohandu et al., 1977), selects the most important variables that describe the main part of variation of the data set and reduces the dimensions without creating new ones.

- For every column of a correlation matrix C of n variables, a measure of influence S_j is calculated.
- $S_j = \frac{\sum_{i=1}^{n} c_{ij}^2}{c_{jj}}$
- The maximum of those measures S^k =max S_j identifies the most important among them. That is the number, on average, of the most important variable in the system. Superscript k indicates the number of the current iteration.

- The correlation coefficients of the maximal variable will be divided by the square root of the diagonal element c_{jj} of the matrix C. The transformed column vector
- $b_1 = \frac{c_j}{\sqrt{c_{jj}}}$
- is the first vector of the new factor matrix B.
- Find the residual matrix
- $\mathrm{C}^{(1)} = C - b_1 b_1^T$
- Repeat the process r times, where r<=n is the rank of C.

5.4 Preprocessing the student retention data from Tallinn University of Technology

Our database consists of the course declaration data of Tallinn University of Technology (TUT). In the time span of 1997-2010, 1.3 million course attendances have been registeredby a total of 40 000 students. We focus on the graduates of Tallinn School of Economics and Business Administration (TSEBA) of TUT. We want to predict university graduation at the **end of the freshman year**, i.e. the point when students have taken the **first 10 courses of their study program**. All in all, there are 928 such students, 633 of whom will be selected by us. The criterion is that the student started their studies at least four years prior to 2010, so that there would have been sufficient time for graduation. We have 425 students who successfully graduated, and 208 who did not. The ratio of students graduating was roughly 2/3 constantly throughout the years. When we tried to use the data of students, who started their studies in 2007 (three years prior to our data collection period) and would have had to graduate in three years, the ratio was substantially different. Thus, such students were not included, as it was decided that the insufficient study period would distort the results. In contrast to most studies on student retention (e.g. (Gray et al., 2013a)), we are not focusing on generating the prediction after the first semester of studies, but rather on determining the fate of students after they have completed their freshman year. Allowing faculty members to intervene at this point will let one follow the dynamics of the progress of students through the university longitudinally. Considering that university management will always have a cost-benefit analysis component to it, it is necessary to ensure that students who have shown to have “higher value” to the society not only begin the university studies well, but also stay on course

Table 7 Feature salience detection using the peeling method

Data item	Course6	Age	Course4	Course10	Estonian/Russian	Man/Woman	Course1
Measure of influence	23,83	15,04	11,96	9,90	8,73	7,52	6,45
Decrease after removal	**36,87%**	**12,95%**	**8,63%**	**4,91%**	**5,08%**	**4,49%**	**4,34%**
Cumulative decrease	36,87%	49,83%	58,46%	63,36%	68,44%	72,93%	77,27%

Data item	High school	Course7	Course5	Course8	Course9	Course3	Full-time/Part-time
Measure of influence	5,42	4,43	3,08	2,28	1,56	0,92	0,43
Decrease after removal	**4,14%**	**5,66%**	**3,39%**	**2,99%**	**2,71%**	**2,07%**	**1,78%**
Cumulative decrease	81,41%	87,06%	90,45%	93,44%	96,15%	98,22%	100,00%

In addition to the attendance data, we know whether the student is **a male or a female**, what the person's **age** at the time of accession to the university is, what the **mother tongue** is, whether they are **a full-time or a part-time** student, and **how good the previous educational institution (**high school**) is.** The mother tongue is recorded to be either Estonian or Russian, corresponding to our demographic division. Every year a leading weekly, Eesti Ekspress, publishes high school ratings that show the average performance of recent high school graduates in the national graduation examinations. We use the ratings of schools from 2008 as they are comparatively constant.

All the data was normalized and discretized to integer values 0..5, so that it would similarly vary in the same range. Course attendance data naturally followed this set of bins. The level of the previous educational institution, according to the ratings, was divided into 6 equal width bins and discretized. The same was done with the age at the time of accession. The binary variables – mother tongue (Estonian/Russian), sex, full time/part time studenthood – were coded as 0/5.

When we analyzed the distributions of the 10 courses taken, it was found that one of the courses resulted in all the students passing. This course was removed from the data because a unit column would have resulted in unsuccessful optimizations. Furthermore, two courses had only 14 failed results. The probability that we would draw 300 times out of 400 without ever hitting any of the 9 (9=14*(400/633)) failed ones was calculated to be virtually nonexistent. Thus this course was not removed.

5.5 Analyzing the histograms of bootstrapping the feature salience-enabled optimization

Initially, feature salience detection was undertaken. Table 7 shows the decrease in the measure of influence after the removal of successive columns as they were identified. Due to the fact that our data originally had 14 dimensions, we decided to compare this to retaining about 80% of variation, condensed to the 7-9 most important dimensions of datawhich is still a reasonable choice for discriminant analysis. We also compared results with two random selections of variables, where we removed 6 variables completely in random and where we contrastingly removed the 6 *most important* variables as assessed by the peeling method.

As is evident from the results, the fate of a student is most influenced/determined by Course6 (this was a quantitative freshman class) and the age at the time of accession. This is in concordance with (Gray et al., 2013b), who takes special measures to encode the age of the student. Notably, the weakest distinguishing indicators are Courses 9 and 3 (both with only 14 negative results), and whether the student studies full- or part-time.

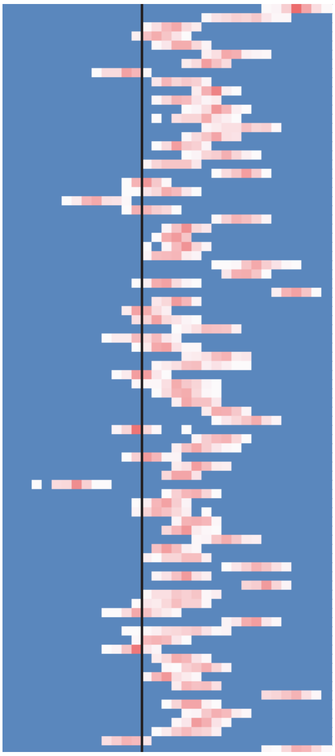
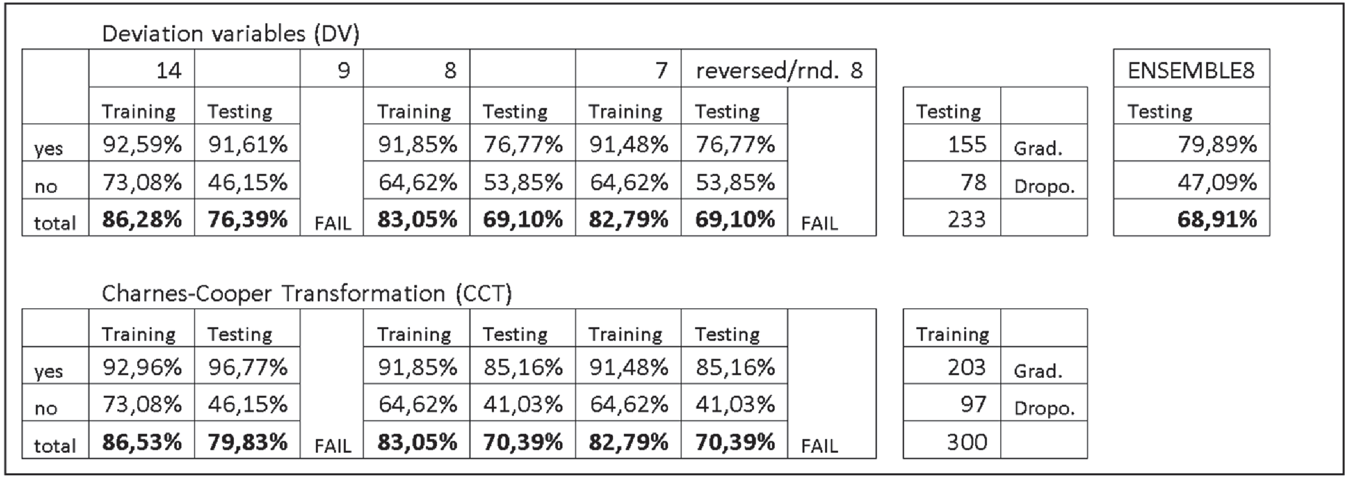


Figure 20 Histograms resulting from bootstrapping. The black line marks the dividing hyperplane. The color marks the frequency of values in that range. There are 33 bins for both families: -0.08..0.12, step 0.00625 for CCT and -65..95, step 5 for DV

The data on 633 students was divided into a training/validation set of 400 and a testing set of 233. The training/validation data set was further partitioned into a training set of 300 students and a validation set of 100 students. This was done repeatedly by applying bootstrapping. Data was partitioned between training, validation and testing, by retaining the same proportion of graduating and dropping-out students as in the entire data set; expect for a brief, unsuccessful experiment. During this experiment we trained an unbalanced data set, which consisted of an equal number of graduates and drop-outs. The results were inferior to all our other predictions. Therefore, we chose to persevere with proportional distributions.

Table 8 Results of training/validation and testing – correct answers. A number of variables varies between 14 and feature salience selected subset, as well as a random selection



Deviation variables (DV)

	14		9	8		7	reversed/rnd. 8	
	Training	Testing		Training	Testing	Training	Testing	
yes	92,59%	91,61%		91,85%	76,77%	91,48%	76,77%	
no	73,08%	46,15%		64,62%	53,85%	64,62%	53,85%	
total	**86,28%**	**76,39%**	FAIL	**83,05%**	**69,10%**	**82,79%**	**69,10%**	FAIL

Testing	
155	Grad.
78	Dropo.
233	

ENSEMBLE8
Testing
79,89%
47,09%
68,91%

Charnes-Cooper Transformation (CCT)

	Training	Testing		Training	Testing	Training	Testing	
yes	92,96%	96,77%		91,85%	85,16%	91,48%	85,16%	
no	73,08%	46,15%		64,62%	41,03%	64,62%	41,03%	
total	**86,53%**	**79,83%**	FAIL	**83,05%**	**70,39%**	**82,79%**	**70,39%**	FAIL

Training	
203	Grad.
97	Dropo.
300	

In order to perform bootstrapping, 300 students were selected as training data and our models of Charnes-Cooper transformation and deviation variables were optimized. The remaining 100 students from the training/validation set were used for validation. The same hyperplanes were used on the testing set, creating predictions about the students whose fate we wanted to know.

Bootstrapping was performed 1000 times (2000 opimizations altogether for the two models), and the results (both for training/validation and testing) were recorded (see Table 8). This resulted in histograms (see Figure 20), which reflect the behavior of the student in relation to the discriminatory hyperplane family. For the test set these were calculated based on the 1000 distances from the hyperplane, whereas in the training/validation set each student had two histograms – one for instances of training and one for instances of validation selection. Yet, even the validation histograms were calculated based on at least 200 values, which was considered sufficient (this student was selected 800 times for training). For training the minimum number of selections for a student was 720 (this student was selected 280 times for validation).

Decisions were undertaken based on the average of the histograms. The use of histograms was justified because the decisions based on the average were better than the average of all the individual decisions. Even more importantly, decisions taken based on individual iterations varied significantly, whereas the average-based decisions guaranteed a certain level of accuracy.

As is usually the case (see Table 8), the results are better for training than for testing. The validation percentages are not explicitly put forth because they coincided almost exactly with the training percentages.

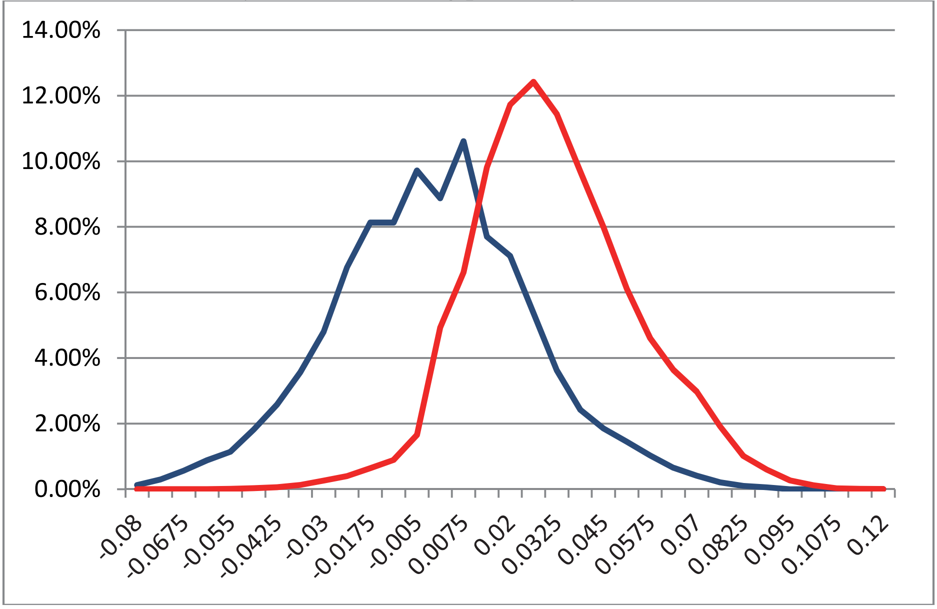


Figure 21 Average histograms of dropouts and graduates, CCT

The highest total testing result was 79.83% for CCT with all 14 variables included. On the other hand, the DV approach that included 8 variables, as prescribed by peeling, gave the highest true-negative test score of 53.85% – something we were interested in as regards dropout forecasting. More generally we found (on the test set) that CCT tends to predict the true-positives better, whereas DV does better on true-negatives.

Therefore, we also tried an ensemble approach on 8 variables. In this case we did not average over the 2000 values, as the histogram bins were different, but rather used the counts (number of predictions to either side of the hyperplane). (The correlation between the count based and average based decisions was 1 for individual model predictions.) As is evident from Table 8, the ensemble-approach did find a compromise between the CCT and DV methods. It remains as future work to combine the very best cases, e.g. the 14 variable CCT with the 8-variable DV, where the number of variables differs but joint decisions can be taken.

There were some combinations of variables, where the model performed dismally, e.g. by predicting all but two true-negatives to graduate. One such case occured when 9 more influencial variables were included, whereas the other was a comparison where the 6 *most important* variables were removed. The optimization with completely random variables selected performed worse than the comparative peeling enabled and fully inclusive approaches.

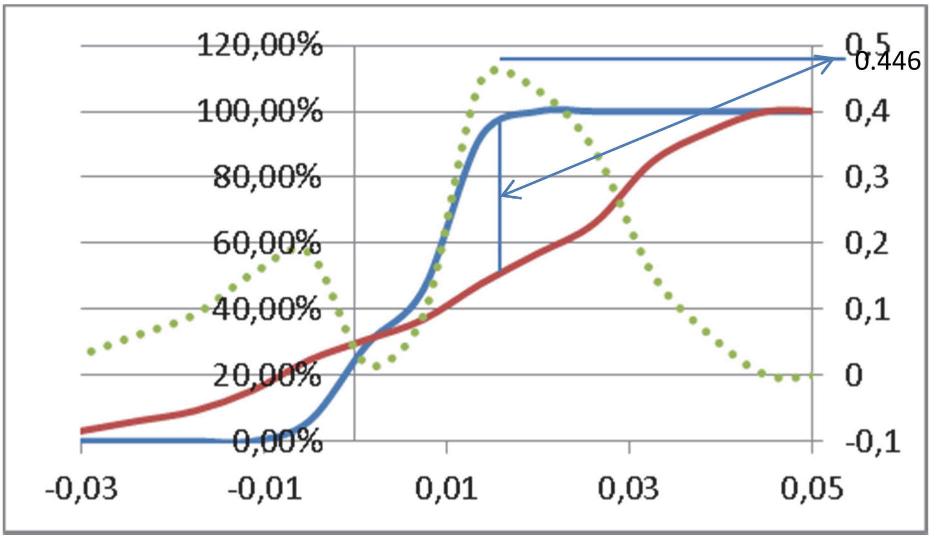


Figure 22 Calculating the Kolmogorov-Smirnov test, with our sample sizes at level 95%, the critical difference allowed is 24%.

Finally, we used a five-fold Kolmogorov-Smirnov (KS) test to generate decisions for the staff to act on. We calculated the average histograms of dropouts and graduates (see Figure 21). The Kolmogorov-Smirnov test measures the maximum distance between cumulative distribution functions and tests whether they have sufficient similarity. The maximum distance allowed depends on the sample sizes and the level of probability. In our case the chosen level(95%)

resulted in the maximum allowed difference of 24% (see Figure 22 for an example).

The following five KS tests were undertaken: whether the student is sufficiently similar to the “profile” of the (average) graduate; whether he is sufficiently similar to the “profile” of the (average) dropout; whether his distribution function is to the “right” of the one of the graduate (very certain to graduate); whether his distribution function is to the “left” of the one of the dropout (very certain to drop out); or whether he is “between” the graduate and dropout profiles, i.e. “uncertain”.

During the development we also tested our models by changing their free parameters. For example, we changed the y_0 parameter of the Charnes-Cooper model, yet without much success, except that we checked some values which were definitely out of range. Thus, 0.01 was selected as the most obvious candidate. For the deviation variable model α^* was selected as 0.5 and β^* as 30 000, in the same way.

While data preprocessing was performed in R; Excel, its VBA and Frontline Solver Platform was used for all other tasks. One bootstrapping execution took about 200 minutes.

5.6 Concluding discussion

We have created an ensemble method for predicting student retention. After preprocessing, normalizing and discretizing the data on the freshmen courses and the demographic variables to 0..5 integer values, we applied the peeling method for feature salience detection. A quantitative course and the age of the student at accession were the first critical dimensions identified. An example of data possibly omitted (amongst the three least influential according to peeling) were the two courses which resulted in only 14 failed exams for our 633 students. Thereby we preferred to select either the full 14 dimensional data set or retain 80% of variation in the data by omitting the 6 less influential dimensions.

We thereafter partitioned the students to a 400 training/validation and a 233 testing set. Training/validation was undertaken by performing bootstrapping, whereby 300 students were used for training and 100 for validation. Bootstrapping was performed on two models – a classical Deviation Variables approach to linear programing discriminant analysis and our application of Data Envelopment Analysis (Charnes-Cooper transformation) in the context.

1000 bootstrapping iterations were undertaken and histograms about the way a student relates to the discriminatory hyperplane family were recorded. Decisions were taken based on the average of the histogram, which guaranteed more stability in testing results, as well as a higher result than the average of 1000 individual testings.

The DV model with 8 variables predicted the true-negatives that we were searching for in the best way. The CCT model with all 14 variables included had the best overall prediction accuracy. We also created an ensemble model with 8 variables that took decisions based on joint information from CCT and DV. This combined the good features of both models (true-positives from CCT and true-negatives from DV), and had a kind of an “average result” from the two. Thus, the model deployment decision should be taken depending on the final goal in mind – whether the true positives or the true negatives are more important.

The models gave the best results after some configuration becausethe free parameters y_0, α* and β* had to be properly chosen. Also, the proportional distribution of graduating/dropping out students had to be chosen over unbalanced data.

Furthermore, histograms to be used by the staff of the university were deployed by applying a five-fold Kolmogorov-Smirnov test, which detailed the level of certainty about the student’s future.

DISCUSSION AND CONCLUSIONS

Conclusion

This thesis has developed methods for the analysis of two university management indicators. On the one hand, it has developed a new indicator – the coopetition score. On the other hand, it has identified factors of and predicted a well-known indicator – student retention. Coopetition was considered in section 1.

While the phenomenon of coopetition is ubiqitous (it exists on the personal, the organizational, and the market level, to name a few), we have concentrated on the organizational level, drawing parallels and **applying business theories** in the educational management context.

We will build our coopetition score on the notions found from literature review, which state that educational institutions (just like multinational corporations) are networked organizations; active in applying mandates and centers; have internal competition and cooperation with the external partners at its fuzzy boundaries; governed by principles of heterarchy and autonomous initiatives.

Considering that on the organizational level the phenomenon has barely been studied, we sought to quantify it with one of the most obvious methods – social network analysis – and thereafter derived a coopetition indicator that can be applied to university management. In that context we discussed mandates (competition) and centers (cooperation) that are evident in the university context.

A department has a mandate to teach its own students, which is its primary goal. This can be measured by the average grade of its graduates. The secondary goal of a department is to be useful for the rest of the university through the breadth of its curriculum. This is measured by the social network analysis indicator, traditional conductance. The final indicator is the equally weighted average of the two measures. Universities face the imperative to actively manage their coopetitive processes.

In our quantification, coopetition is measured through the academic records of the students of a department. We measure the similarity of students by the courses they have taken.

In social network analysis, the communities of students that have been detected mark the departments. Community detection has different optimization criteria that are maximized, and those can either be followed to strict optimality or calculated by heuristics. We use modularity as our objective function. As the number of students is in hundreds, we develop the coopetition indicator by using a heuristic.

Due to the fact that all the values of student similarities can be calculated, we seek to remove the correct portion of those in section 3, while constructing the SNA adjacency matrix. This way, we hope to retain the signal in the data table while discarding the noise, which is represented by the lower similarity value layers.

This time our quantitative process is in the form of supervised learning. We know the departmental specializations of students and seek to remove layers of the adjacency matrix through the distance threshold method, so that the communities are as homogenous as possible. We will, however, use the strictly optimal modularity algorithm which uses linear programming to tackle the NP-hard problem.

We sample 36 students, who belong to two different groups, out of any combinations of our four specializations (Marketing, Management, Accounting, Finance); and optimize modularity. As we aim to find the best distance threshold for removing noise, we optimize matrices of differing sparseness. For our data, a middle value of distance threshold (0.7) is proven to be the best. This is the value at which we have the smallest number of clusters that are as "pure" as possible. As we find out, all the communities (e.g. Marketing) are further divided into two subclusters.

Since we have not used all the data at once, we also seek to verify whether the clusters that we detected were stable, i.e. the same (e.g. Marketing) students always ended up in the same communities (e.g. the same Marketing subcommunity). Subcluster stability is successfully determined by the emerging "checkerboard" pattern for the different combinations of specializations, at the optimal distance threshold value. We now know that the communities have been detected and the coopetition score derived in an optimal way.

Our next concern, in section 2, is the permutations of our coopetitive adjacency matrix. As is suggested by the fact that the strictly optimal solution to the community detection is NP-hard, seriating for useful permutations is equally complex.

We have created a tool for collaborative seriation – Visual Matrix Explorer – for this purpose. This tool enables a streamlined application of nine seriation algorithms, together with the ability of dynamically coloring the data on related permutations and a recursive drilldown feature that reruns the optimizations on a rectangular data subset, which by now has an enchanced number of degrees of freedom.

We make use of this tool on our coopetitive adjacency matrix because we expect a strong block diagonal to appear – a feature that has different, related interpretations in fields such as archeology, cellular manufacturing, social network ananlysis, operations research and anthropology.

While we study the formed block diagonal structure, we drill down on one particular community and, again, find a subdivision within. The BEA algorithm permutation, however, finds an internal order within the two subcommunities, so that the students forming the stronger ties between the two are aligned next to each other on subcommunity borders. The thick chunks that form on the border of the subcommunities are important from the coopetitive point of view (as our indicator uses traditional conductance of communities) because they allow us to identify students pertaining more to the competitive as well as to the cooperative stance.

In section 4, we turn to predicting the student retention indicator**.** We once again find that business theories can be used to inform educational management research. This time our goal is to predict student retention/churn, and in order to do that we turn to retention theories. There already exist theories that have juxtaposed business processes with the retention of students – such as employee turnover for example.

We turn to theories on business networks in order to propose factors such as entry barriers, sunk costs, commitment, learning, formation of routines, mutual adaptation, etc. that have parallels with the student-university relationship. We also find that the determinants and the taxonomy of student churn are in parallel with customer churn and set out to predict student retention in section 5.

In section 5, we again deploy our prevailing method of analysis – linear programming. We identify a number of input factors that provide data for a discriminant analysis problem. These include the first 10 courses because we are predicting retention after the freshman year. Furthermore, a number of demographic variables enter the equation – age at accession, sex, mother tongue, high school level, full time/part time studenthood. The data is preprocessed and discretized.

Our method will deploy two multiple criteria optimization methods, one using deviation variables and the other deploying Charnes-Cooper transformation to linearize a fractional programming problem. The decision is not taken after a single execution, instead bootstrapping is used by further dividing the training sample into training and validation and repeatedly sampling them at random.

Thereafter histograms are constructed that relate the discriminatory hyperplane family to the data point. The decision is taken based on the simple average of the histogram which results in better predictive power and less variability than would have been possible using single optimizations.

Since section 4 had identified a need to link the customer retention effort to its "value", we begin to address this call by taking a fivefold Kolmogorov-Smirnov test on the histograms. Students are categorized, with those in the "uncertain" category (those most narrowly receiving a negative prediction) getting the biggest effort, when all other things are equal.

Our final step is combining and varying the models. We use feature salience detection in the form of peeling, to select a subset of variables – ones that are related to the biggest proportion of variation in the data. In addition to removing some input variables, we form an ensemble method that combines the DV and CCT methods.

The results show that CCT does best in the overall prediction accuracy, whereas the DV method with removed variables is more successful in estimating true-negatives, which we seek. An ensemble method performs at a level that is the average of the two.

As was mentioned above, the relationship that a student has with the university can be explained as akin to that developing between business partners. Its essential characteristics are an increasing commitment to and investment in one another, which is why it is important to lessen churn **in the beginning of the studies**, before the sunk costs have been formed. This is what we set out to do. We also investigate how **part-time studenthood influences the study outcome** as there are considerable entry barriers to overcome, but this negative relationship with the end result is not borne out by the feature salience detection.

Similarly to business networks, the relationship lies somewhere between the hierarchical fiat and the arm's length transaction, which is why it is important to consider the adaptation, learning, knowledge transfer and formation of routines that takes place. Thus, as empirical investigation also confirms **the quantitative introductory courses to be the main hurdle** whenentering the studies, there might be merit to the idea developed in verbal analysis for postponing those courses until second year by bringing the more specialization specific courses to freshman year.

The theoretical study also identifies **age** as an important factor when successfully entering studenthood as the personal changes of becoming an adult are overcome. This is confirmed by the empirical investigation.

Thus, we have successfully developed methods for analyzing two university management indicators – coopetition and student retention. We have shown the state of affairs in educational institutions to be similar to that in the business venue. Figure 25 summarizes the work by showing how business processes inform educational management. The data are utilized for developing indicators of educational management through educational data mining. It directly benefits both the students (retention rate) and the departments of the university (coopetition score).

Contributions and answers to research questions

The main contributions of the thesis are summarized on Figure 23 and Figure 24 and these correspond to the major research subquestions.

Figure 23 sums up the main contributions of our coopetition study. Business management literature has helped us understand educational organizational coopetition in a wider context. A number of MNC management related traits have direct parallels in the university context, as stated in the beginning of the conclusion. Most importantly, we have identified the concepts mandate and center as the basis for our coopetition indicator. We have developed the indicator by using SNA, basing it on the average grades and traditional conductance, and by applying modularity community detection methods. However, measuring our indicator has to be facilited in the best possible way. We have retained the signal and discarded the noise, using the distance threshold method for this. As for the second pillar that supports our coopetition study, we sought to learn about the internal structure of the communities. Therefore, we developed the seriation tool VME and identified the internal order of the leaves on the borders of the two subcommunities during a drilldown operation on the modularity permutation. We have sought to ***apply the data scientist approach to management*** *in our coopetition study. Therefore,* ***we have combined the analysis of the problem regarding coopetition with a number of algorithmic improvements, which lead us to an interdisciplinary solution****. The data scientist approach, however, cuts both ways in the sense that all the developments can also be separately applied in their respective domains. Therefore,* ***our coopetition indicator can in turn also be reapplied in the business management context.*** *Additionally, the topics of* *seriation and retaining the signal over the noise can be put to domain specific use, all of which is depicted on Figure 23.*

Figure 24 describes the contributions of our retention prediction effort. We once again apply business theories to educational management, showing its usefulness. As we identify important factors, we see that it makes sense to deal with retention predominantly in freshman year. We also seek to differentiate the part-time students and correctly predict the introductory quantitative courses of the technology school as well as the age at accession, all important determinants of churn. Subsequently, we identify input variables for our retention prediction. Thereafter, we make the prediction by using the deviation variable and Charnes-Cooper transformation ensemble linear programming discriminant analysis with bootstrapping, feature salience detection and histogram based decision making. As regards our retention effort, the ensuing numerical method can also be applied in other contexts and is independent of the current data.

The takeaway points, which answer the six minor research questions, are:

- **Research about the management of multinational corporations provides insights for coopetition research by outlining the logic behind the centers and mandates.**
- **Business reseach provides insights for educational management, as regards both organizational coopetition and student retention. In the context of retention, the theories that consider churn and business network relationships are viable for consideration, and the implications include the consideration of customer lifetime value, the taxonomy of churn and the business network relationship factors (e.g. during this consideration, the age at accession can be identified as an input factor for predicting retention).**
- **Mandates are measured by the average grades and centers by the SNA metric traditional conductance from a modularity heuristic on a student similarity matrix, and the two are combined into an indicator.**
- **By varying the sparsity of the matrix, we retain signal and remove noise from the adjacency matrix, using the distance threshold method. There is an optimal amount of layers to be removed, as measured by strictly optimal modularity.**
- **Visual Matrix Explorer allows us to further explore the communities, exposing the dynamic internal structure of the communities detected while using the drilldown feature.**
- **We classify student retention data by using the deviation variable and Charnes-Cooper transformation ensemble linear programming**

discriminant analysis with bootstrapping, feature salience detection and histogram based decision making.

And to sum it up, we can state, as regards the main reseach question (that was answered throughout the thesis) (see Figure 25):

- **Since the business processes are becoming increasingly dynamic, networked and multilateral, we have shown that these can inform the educational actors by providing useful managerial tools considering that the central imperative is still to influence the institutional outcome.**

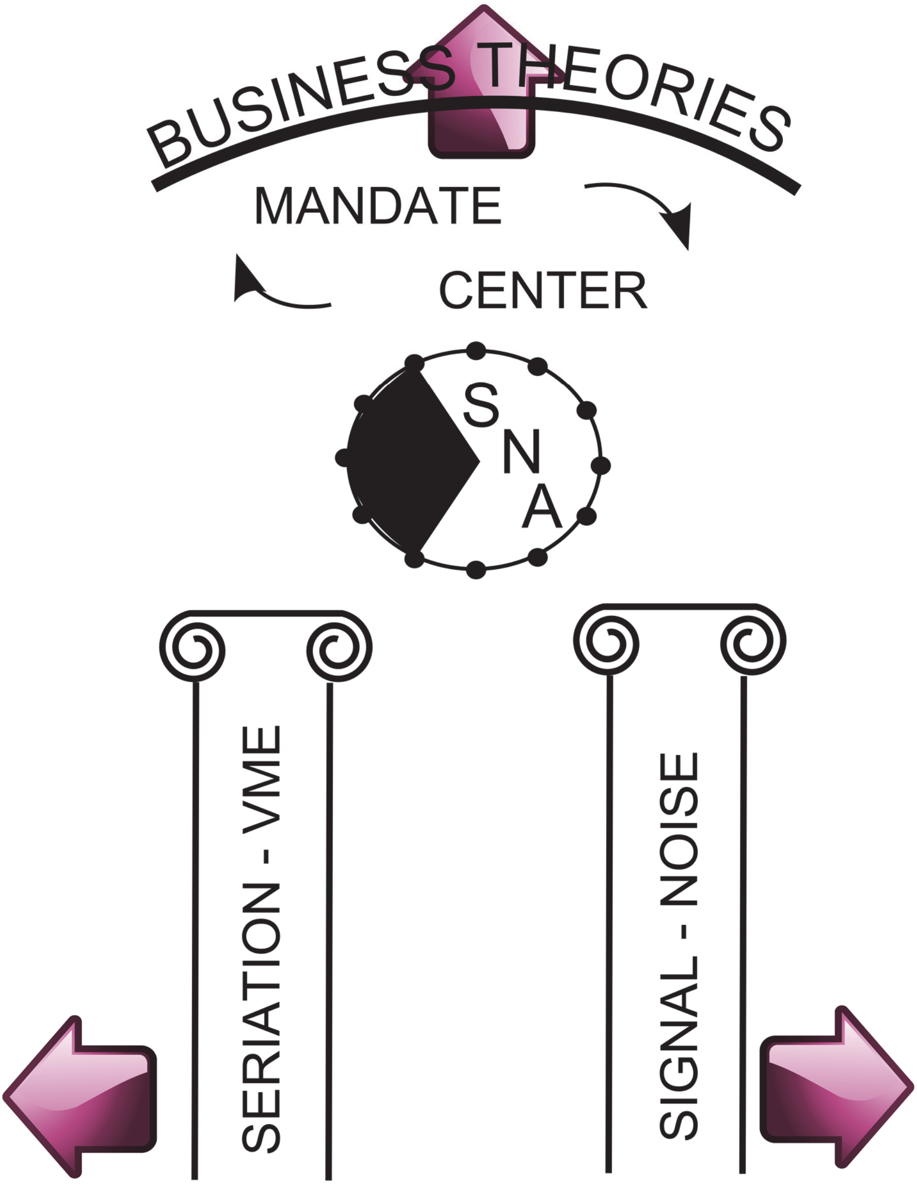


Figure 23 Results of the coopetition study and applicability of results in other contexts

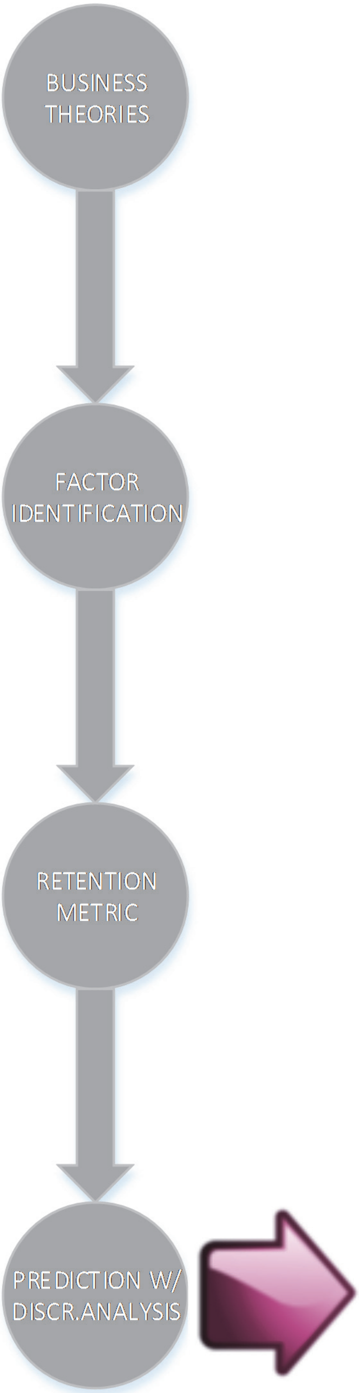


Figure 24 Results of the churn study and applicability of results in other contexts

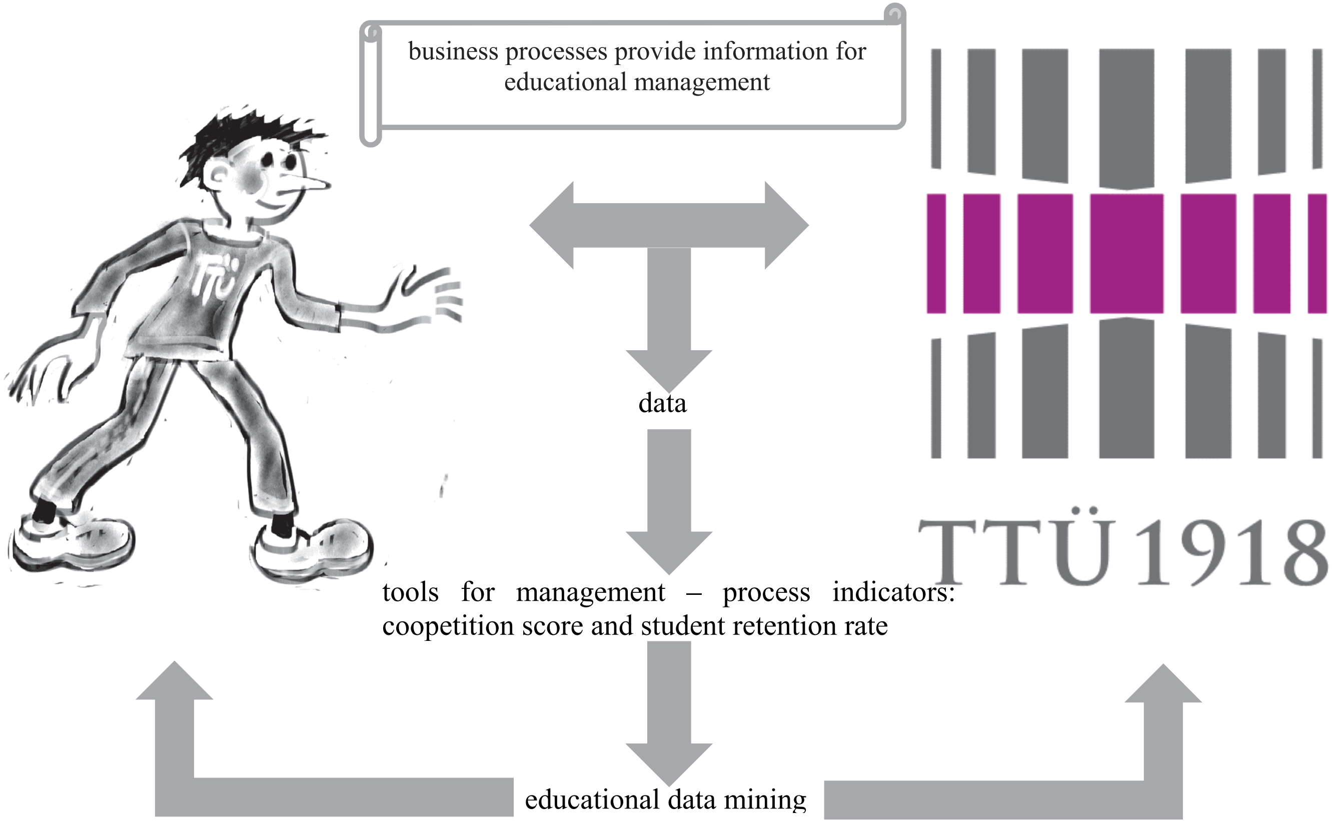

business processes provide information for educational management
data
tools for management – process indicators: coopetition score and student retention rate
educational data mining
TTÜ 1918
TTÜ

Figure 25 Factors and tools influencing the student-university relationship

The student-university relationship – a third viewpoint, that of a faculty member

By quantifying university management indicators in this thesis, we have considered the student-university relationship, and taken the viewpoint of a university (department), which is in coopetition for students, and that of a student, who needs not to churn. In this subsection, however, we are going to take the viewpoint of a faculty member, whose goal is to publish and teach classes related to his research area. The autor of this thesis has been doing the latter as well. Thus, this subsection is going to put forth an overview of the ways in which the student-university relationship has been influenced.

The author, who did his Bachelor's thesis in business administration, afterwards specialized in informatics. The goal of the author, as a lecturer, has been to develop a quantitative curriculum module that can be taught by the Faculty of Informatics in the Tallinn School of Economics and Business Administration as well as in the Department of Logistics in the Faculty of Civil Engineering.

As of this autumn, the author is coordinating the teaching of a module that begins with a course "excel43.com" (Informatics I – semester 1), a quantitative introduction to business/logistics studies, which introduces students to many topics considered during their Bachelor's studies. This course provides them with the core spreadsheet way of thinking, is an alternative to the way quantitative matters are introduced following the AACSB mandate, and is an original development.

This is followed by a course "Optimization modelling with spreadsheets" (Informatics II – semester 4), which considers a wide range of optimization/sensitivity analysis problems that have a background in business and logistics. This course uses a subset of a traditional Decision Analysis textbook and deploys the standard Solver included with Excel.

While programming has traditionally been taught to all undergraduate students of the different faculties of our tech school as an Informatics II course, this strand is now only an elective for business/logistics, possibly replacing the Informatics II course. Students who are more interested in deepening their quantitative skills are the ones who usually take this class. However, it is recommended to all students who plan to enter Master's studies, as "they can concentrate on decision analysis later on as well and as no one wants to study programming while already at the Master's level".

The set of problems tackled in the Algorithms for Business and Logistics course include topics such as programming net present value; internal rate of return; a travelling salesman's problem with simulated annealing heuristic; bucket brigades algorithm for putting out fires; a graphical solution to self-organizing maps to name a few, and is in the phase of active development.

This course is followed by a semester 7 (Master's program) course "Decision Analysis" for business and logistics, which has two strands. The standard textbook is followed, for those who took programming during Informatics II of their earlier studies. Those who have already taken the optimization block are encouraged to take the version of the class that includes two new topics – a combination of Solver and VBA programming from one textbook; as well as a combination of various Excel Add-Ins with VBA (including MapPoint, RiskOptimizer, XLStat) from another textbook.

The final course is from semester 9 (Master's program) and is titled "Revenue Management for Business and Logistics" – a hot topic that integrates microeconomics, statistics and optimization. The topics taught include market segmentation, arbitrage, cannibalization, market segmentation with supply constraints, variable pricing, capacity allocation, network management, overbooking, and markdown management. The most widely used textbook in the field is studied in this course.

Two courses – Informatics I and Informatics II – are fully youtubecast and together with colleagues we have introduced recent concepts, e.g. the use of Scratch and computational thinking, in programming.

As far as business and logistics students are concerned, the courses that have been developed complement the data mining courses that are read in the advanced levels of study throughout our university by my supervisor, Innar Liiv.

The development of these courses complements the departmental coopetition and student retention research by taking the third viewpoint, at the personal level, of a faculty member.

Further research

One further research direction would be the different manifestations of coopetition. As intraorganizational coopetition has not been extensively studied, it could be further investigated. Since research does not have to exclusively pertain to educational data mining, one of the most interesting research venues would be the multinational corporations themselves. There, the coopetitive processes are definitely at their most intensive, while the depth and breadth of the information provided is at its most extensive. In the case of a cooperation agreement with a big multinational, one would be provided with a unique set of data for studying organizational coopetition.

In the educational data mining context, coopetition needs further scrutiny as well. As many students reach doctoral studies by the end of their relationship with the university, coopetition can be stated to move from those for business partners to those for employees. Information and Communication Technology is the biggest amongst the faculties of TUT and provides a curious case, as it is the one creating several possibilities for interdisciplinary linkages. Our department alone has ongoing projects of cooperation with bioinformatics, medicine, business, logistics, law, and graduating doctoral students have moved to many other organizational departments from here.This creates a need to investigate coopetition for employees.

We can say that there are three areas of activity, which are nowadays competing for the attention of faculty members – research, teaching and industry cooperation. TUT presents an interesting case for studying this, too, because there is a top level organizational unit, MEKTORY that exists for the sole purpose of promoting industry cooperation. As a result, a faculty member, who engages in all three, will have two competing bosses to respond to, as well as different remuneration schemes to optimize his behavior for. Also, coopetition between departments and MEKTORY needs to be studied.

As regards the methods of studying educational coopetition, we have applied network science, whereas the other very interesting method – agent oriented simulation – has not yet been applied.

As for educational data mining, business theories could further be applied in educational context. For example, there seems to be much merit in following the popular trend of linking different sources of data. This way, student data could be tracked from the E-School (high school grades), through the university studies, all the way to the job market (including salary data) in an attempt to understand the coopetition that takes place between the students.

When considering matrices, e.g. the student similarity matrix, one further venue of research would be trying to expand the application of the use of Kolmogorov complexity as a tool for measuring the order attained in the permuted matrix. Right now such compression operations are undertaken, following the z-curve, but other possibilities can be considered.

The two data tables that have been used in this thesis (a table with the demographical information on students and the table regarding course declarations) are a seemingly endless source for educational data mining research. To take one example, the paths of students on their set of courses during the study program can be tracked in order to determine who is better in how many courses (measuring an aspect of their coopetition), and asymmetric clustering can be used on the resultant table.

REFERENCES

1. Ackerman, R. & Schibrowski, J. (2008). ‘A business marketing strategy applied to student retention: a higher education initiative’. J. of Col. Stud. Retent., 9:307-336.
2. Adersson, M. & Holmström, C. (2000). ‘The dilemma of developing a centre of excellence – a case study of ABBGeneration’. In U. Holm & T. Pedersen, editors, The emergence and impact of MNC centres of excellence. 271 pg. Macmillan.
3. Allen, R., & Burgess, S. (2013). ‘Evaluating the provision of school performance information for school choice’, Economics of Education Review, pp. 175-190.
4. Andersson, M. and C. Holmström (2000). 'The Dilemma of Developing a Centre of Excellence-a Case-Study of ABB Generation', Diva-Portal.org.
5. Astin, A.W. (1997). ‘What matters in college: Four critical years revisited’. San Fransisco: Jossey-Bass.
6. Bajgier, Steve M., and Arthur V. Hill. ‘An experimental comparison of statistical and linear programming approaches to the discriminant problem.’ Decision Sciences 13, no. 4 (1982): 604-618.
7. Bal, H., Orkcu, H., Celebioglu, S. (2006). ‘An experimental comparison of the new goal programming and the linear programming approaches in the two-group discriminant problems’. Computers & Industrial Engineering.50, 296-311.
8. Bazaraa, M. S., J. J. Jarvis, H. D. Sherali and M. S. Bazaraa (1990). Linear Programming and Network Flows. Vol. 2. Wiley Online Library.Birkinshaw, J. (1995). 'Business development initiatives of multinational subsidiaries in Canada', Gouvernement du Canada-Industrial Organization (Gouvernement du Canada-Industry Canada).
9. Bean, J.P. & Eaton, S.B. (2000). ‘A psychological model of college student retention’. In Braxton, J.M., editor, Reworking the student departure puzzle, 48-61. Nashville: Vanderbilt University Press.
10. Bean, J.P. & Metzner, B. S. (1985). ‘A conceptual model of nontraditional student attrition’. Review of Educational Research 55: 485-540.
11. Bean, J.P. (1980). ‘Dropouts and turnover: The synthesis and test of a casual model of student attrition’. Research in Higher Education 12: 155-187.
12. Berger, J.B. (2000). ‘Optimizing capital, social reproduction, and undergraduate persistence: a sociological perspective’. In Braxton, J.M., editor, Reworking the student departure puzzle, 95-126. Nashville: Vanderbilt University Press.
13. Bertin, J. (1981). ‘Graphics and graphic information processing’. Walter de Gruyter.
14. Birkinshaw, J. & Hood, N. (1998). ‘Multinational corporate evolution and subsidiary development’. 392 pg. London: Macmillan.

15. Birkinshaw, J. (1997). 'Entrepreneurship in multinational corporations: The characteristics of subsidiary initiatives.', Strategic Management Journal 18, pp. 207-229.
16. Birkinshaw, J. (2000). ‘Entrepreneurship in the Global Firm’. 165 pg. Mcmillan.
17. Birkinshaw, J. (2001). Entrepreneurship in the Global Firm. Taylor & Francis.
18. Blondel, V. D., J. L. Guillaume, R. Lambiotte, and E. Lefebvre (2008). 'Fast unfolding of communities in large networks', Journal of Statistical Mechanics: Theory and Experiment, P10008.
19. Brandenburger, A. and B. Nalebuff (1998). Co-opetition. Crown Business.
20. Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., & Wagner, D. (2008). ‘On modularity clustering’. Knowledge and Data Engineering, IEEE Transactions on, 20(2), 172-188.
21. Braxton, J. M. (2003). ‘Persistence as an essential gateway to student success’. In Komives, S. & Woodard, D., editors, Student services: a handbook for the profession. 4th edition: 317-335. San Francisco: Jossey-Bass.
22. Brizendine, L. (2006). The Female Brain. Broadway.
23. Buckley, P. J., and M. C. Casson, (1998). 'Models of the multinational enterprise', Journal of International Business Studies, pp. 21-44.
24. Burbidge, J. L. (1971). ‘Production flow analysis’. Production Engineer, 50(4.5), 139-152.
25. Burgelman, R. A. (1983). 'A model of the interaction of strategic behavior, corporate context, and the concept of strategy', Academy of Management Review, pp. 61-70.
26. Burgelman, R. A. (1994). 'Fading memories: A process theory of strategic business exit in dynamic environments' Administrative Science Quarterly, pp. 24-56.
27. Burgelman, Robert A. (1994). ‘Fading memories: A process theory of strategic business exit in dynamic environments’. Admin. Science Quarterly, 1(39): 24-57.
28. Carroll, J. D., & Arabie, P. (1983). ‘INDCLUS: An individual differences generalization of the ADCLUS model and the MAPCLUS algorithm’. Psychometrika, 48(2), 157-169.
29. Caves, R. E., & Porter, M. E. (1977). ‘From Entry Barriers to Mobility Barriers: Conjectural Decisions and Contrived Deterrence to New Competition’. The Quarterly Journal of Economics, 241-261.
30. Chandrasekharan, M. P., & Rajagopalan, R. (1987). ‘ZODIAC—an algorithm for concurrent formation of part-families and machine-cells’. International Journal of Production Research, 25(6), 835-850.
31. Chandrasekharan, M., & Rajagopalan, R. (1986). ‘MODROC: an extension of rank order clustering for group technology’. International Journal of Production Research, 24(5), 1221-1233.

32. Chen, J., Zaïane, O. R., & Goebel, R. (2009, April). 'Detecting Communities in Social Networks Using Max-Min Modularity'. In SDM (Vol. 3, No. 1, pp. 20-24).

33. Chen, M. J. (2008).'Reconceptualizing the Competition-Cooperation Relationship', Journal of Management Inquiry, 17, pp. 288-304.

34. Chung, Y., & Kusiak, A. (1991). 'GT/ART: Using Neural Networks To Form Machine Cells'. Manufacturing Review 1991, 4:293–301.

35. Collier, D. and R. Adcock (1999). 'Democracy and dichotomies: A pragmatic approach to choices about concepts.', Annual Review of Political Science, 2, pp. 537-565.

36. Czekanowski, J. (1909). 'Zur differentialdiagnose der Neandertalgruppe'. Friedr. Vieweg & Sohn.

37. Danon, L., Diaz-Guilera, A., Duch, J., & Arenas, A. (2005). 'Comparing community structure identification'. Journal of Statistical Mechanics: Theory and Experiment, 2005(09), P09008.

38. D'Cruz, J. (1986). 'Strategic management of subsidiaries'. In H. Etemad & L.S. Dulude, editors, Managing the multinational subsidiary. 224 pg. Croom Helm.

39. Devarics, C. & Roach, R. (2000). 'Fortifying the federal presence in retention'. Black Issues in Higher Education 17(3):20-25

40. Economist (2002). 'Life, the universe and everything'. www.economist.com

41. Economist (2006). 'The birth of philanthrocapitalism'. www.economist.com

42. Eisenhardt, K. M. and D. C. Galunic (2000). 'Coevolving: At last, a way to make synergies work', Harvard Business Review, 78, pp. 91-101.

43. Erenguc, S., Koehler, G. (1990). 'Survey of mathematical programming models and experimental results for linear discriminant analysis'. Managerial and Decision Economics.11, 215-225.

44. Ferber, M. A. and J. A. Nelson (1993). Beyond Economic Man: Feminist Theory and Economics. University of Chicago Press.

45. Forsgren, M. and C. Pahlberg (1992). 'Subsidiary influence and autonomy in international firms', Scandinavian International Business Review, 1, pp. 41-51.

46. Forsgren, M. and T. Pedersen (1998). 'Centres of excellence in multinational companies: the case of Denmark.' in The Multinational Corporate Evolution and Subsidiary Development, London and New York: MacMillan, pp. 141--161.

47. Forsgren, Mats & Holm, Ulf & Johanson, Jan (2005). 'Managing the embedded multinational: A business network view'. 227 pg. UK: Edward Elgar.

48. Forsgren, Mats & Johansson, Jan (1992). 'Managing in international multi-center firms'. In Mats Forsgren & Jan Johansson Managing networks in international business. 250 pg. Philadelphia: Gordon & Beach.

49. Fortunato, S. (2010). 'Community detection in graphs', Physics Reports, 486, pp. 75-174.

50. Freed, N., Glover, F. (1981). ‘Simple but powerful goal programming models for discriminant problems’. Eur. J. of Oper.. Res.7, 44-60.

51. Gale, N., Halperin, W. C., & Costanzo, C. M. (1984). ‘Unclassed matrix shading and optimal ordering in hierarchical cluster analysis’. Journal of Classification, 1(1), 75-92.

52. Galunic, D. C. and K. M. Eisenhardt (1996). 'The evolution of intracorporate domains: Divisional charter losses in high-technology, multidivisional corporations.", Organization Science, pp. 255-282.

53. Ghosal, S. and C. A. Bartlett (1997). The Individualized Corporation: A Fundamentally New Approach to Management. Harper Collins Publishers.

54. Ghoshal, S. and C. A. Bartlett (1990). 'The multinational corporation as an interorganizational network', Academy of Management review, pp. 603-625.

55. Ghoshal, S. and D. E. Westney (2005). Organization Theory and the Multinational Corporation. Palgrave Macmillan.

56. Ghoshal, S., & Bartlett, C. A. (1990). ‘The multinational corporation as an interorganizational network’. Academy of Management Review, 15: 603-625.

57. Glover, F. (1990). ‘Improved Linear Programming Models for Discriminant Analysis’. Decision Sciences.21, 771-785.

58. Gray G, McGuinness, C, Owende, P. (2013a) ‘Investigating the efficacy of algorithmic student modelling in predicting students at risk of failing in tertiary education’. EDM 2013, Memphis, USA

59. Gray G, McGuinness, C, Owende, P. (2013b) ‘An Investigation of Psychometric Measures for Modelling Academic Performance in Tertiary Education’. EDM 2013, Memphis, USA

60. Griffin, A. S., S. A. West and A. Buckling (2004). 'Cooperation and competition in pathogenic bacteria', Nature, 430, pp. 1024-1027.

61. Hao, Y. S., Shih, H. Y., Huang, H. C., & Lin, L. L. (2010). ‘Co-opetition of cooperative and competitive relationship: A network analysis approach’. In Technology Management for Global Economic Growth (PICMET), 2010 Proceedings of PICMET'10: (pp. 1-8). IEEE.

62. Harding, S. G. (2004). The Feminist Standpoint Theory Reader: Intellectual and Political Controversies. Routledge.

63. Hedlund, G. (1986). 'The hypermodern MNC?-a heterarchy?', Human Resource Management, 25, pp. 9-35.

64. Hedlund, G. and D. Rolander (1990). 'Action in heterarchies: new approaches to managing the MNC', in Managing the Global Firm, London: Routledge, pp. 15-46.

65. Hedlund, G. and J. Ridderstrale. 'International development projects: Key to competitiveness, impossible, or mismanaged?', International Studies of Management & Organization, 25, pp. 158-184.

66. Hofstede, G. and G. J. Hofstede (2005). Cultures Consequences. Mc Graw-Hill.

67. Hruschka, D. J. and J. Henrich (2006). 'Friendship, cliquishness, and the emergence of cooperation.', Journal of Theoretical Biology, 239, pp. 1-15.

68. Humes, S. (1993) Managing the Multinational: Confronting the Global-Local Dilemma. Prentice Hall Hemel Hempstead.

69. Jacomy, M., S. Heymann, T. Venturini, and M. Bastian (2011). 'ForceAtlas2, a graph layout algorithm for handy network visualization', Retrieved from Gephi.

70. Kaparthi, S., & Suresh, N. C. (1992). ‘Machine-component cell formation in group technology: a neural network approach’. The International Journal of Production Research, 30(6), 1353-1367.

71. Kianmehr, K. & Alhajj, R. (2009). ‘Calling communities analysis and identification using machine learning’. Expert Systems with Applications 36:6218-6226.

72. King, J. R., & Nakornchai, V. (1982). ‘Machine-component group formation in group technology: review and extension’. The International Journal of Production Research, 20(2), 117-133.

73. Kirjavainen, T., & Loikkanent, H. A. (1998). Efficiency differences of Finnish senior secondary schools: an application of DEA and Tobit analysis. Economics of Education Review, 17(4), 377-394.

74. Kirstuks, D. (1999). 'Subsidiary Evolution in Estonia and Latvia: A Case Study of ABB', PhD thesis.

75. Kirt, T. (2002). ‘Combined method to visualize and reduce dimensionality of the financial data sets’. In H.M. Haav, & A. Kalja (Eds.), Proceedings of the fifth international Baltic conference BalticDB&IS 2002 (Vol.2, pp. 255–262). Tallinn: Institute of Cybernetics at Tallinn University of Technology. - Tallinn.

76. Klemperer, Paul (1987). ‘The competitiveness of markets with switching costs’. Rand Journal of Economics, 18 (1): 138-150

77. Koehler, G.. (1990). ‘Considerations for mathematical programming models in discriminant analysis’. Managerial and Decision Economics.11, 227-234.

78. Kou, G., Liu, X., Peng, Y., Shi, Y., Wise, M., Xu, W. (2003). ‘Multiple criteria linear programming approach to data mining’. Optimization Methods and Software.18, 453-473.

79. Krajewski, S., S. Blank, and H.S. Yu (1994). 'North American business integration', Business Quarterly, 58, pp. 55-61.

80. Kuh, G.D. & Love, P.G. (2000). ‘A cultural perspective on student departure’. In Braxton, J.M., editor, Reworking the student departure puzzle, 196-212. Nashville: Vanderbilt University Press.

81. Larson, R. (2009). ‘Education: our most important service sector’. Serv. Sc., 1: I-III.

82. Lenstra, J. K. (1974). ‘Clustering a Data Array and the Traveling-Salesman Problem’. Operations Research, 22(2).

83. Li, G., Ma, Q., Tang, H., Paterson, A. H., & Xu, Y. (2009). ‘QUBIC: a qualitative biclustering algorithm for analyses of gene expression data’. Nucleic acids research, 37(15), e101-e101.

84. Liiv, I. (2010a). ‘Seriation and matrix reordering methods: An historical overview’. Statistical analysis and data mining, 3(2), 70-91.

85. Liiv, I. (2010b). ‘Towards information-theoretic visualization evaluation measure: a practical example for bertin's matrices’. In Proceedings of the 3rd BELIV'10 Workshop: BEyond time and errors: novel evaLuation methods for Information Visualization (pp. 24-28). ACM.

86. Lohk, A., K. Vare, and L. Vohandu (2011). "First steps in checking and comparing Princeton WordNet and Estonian Wordnet', EACL Conference 2012 Proceedings.

87. Lõhmus, Ahto (1974). ‘A set of methods for processing and analysis of information regarding decisions on some aspects of educational process of a university’. Tallinn University of Technology, 87 pg.

88. Mattison, Rob (2005). ‘The telco churn management handbook’. 366 pg. XiT Press

89. McCormick Jr, W. T., Schweitzer, P. J., & White, T. W. (1972). ‘Problem decomposition and data reorganization by a clustering technique’. Operations Research, 20(5), 993-1009.

90. Messner, P. E., & Ruhl, M. L. (1998). ‘Management by fact: a model application of performance indicators by an educational leadership department’. International Journal of Educational Management, 12(1), 23-27.

91. Mironova, O., Amitan, I., Vilipold, J., Saar, M., & Ruutmann, T. (2013). ‘Computer science e-courses for students with different learning styles’. In Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on (pp. 735-738). IEEE.

92. Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007, October). ‘Measurement and analysis of online social networks’. In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement (pp. 29-42). ACM.

93. Moore, K. and J. Birkinshaw (1998). 'Managing knowledge in global service firms: centers of excellence', The Academy of Management Executive, pp. 81-92.

94. Mueller, C., Martin, B., & Lumsdaine, A. (2007). ‘A comparison of vertex ordering algorithms for large graph visualization’. In Visualization, 2007. APVIS'07. 2007 6th International Asia-Pacific Symposium on (pp. 141-148). IEEE.

95. Mullat, I. E. (1976). ‘Extremal subsystems of monotonic systems’. 1. Automation and Remote Control, 37(5), 758-766.

96. Newman, M. E. (2004). ‘Analysis of weighted networks’. Physical Review E, 70(5), 056131.

97. Newman, M. E., & Girvan, M. (2003). ‘Mixing patterns and community structure in networks’. In Statistical mechanics of complex networks (pp. 66-87). Springer Berlin Heidelberg.

98. Niermann, S. (2005). ‘Optimizing the ordering of tables with evolutionary computation’. The American Statistician, 59(1).

99. Peng, T. J. and M. Bourne (2009). 'The Coexistence of Competition and Cooperation between Networks: Implications from Two Taiwanese Healthcare Networks', British Journal of Management,20, pp. 377-400.

100. Peng, T. J., S. Pike, J. C. Yang and G. Roos (2011).'Is Cooperation with Competitors a Good Idea? An Example in Practice', British Journal of Management.

101. Petrie, W. M. F. (1899). ‘Sequences in prehistoric remains’. Harrison and sons.

102. Poole, M. S. and A. H. Van de Ven (1989). 'Using paradox to build management and organization theories', Academy of Management Review, pp. 562-578.

103. Poynter, T. A. and R. E. White (1984). 'The strategies of foreign subsidiaries: responses to organizational slack', International Studies of Management & Organization, 14, pp. 91-106.

104. Qian, Zhiguang & Jiang, Wei & Tsui, Kwok-Leung (2006). ‘Churn detection view via customer profile modeling’. Int. Jour. of Prod. Res., 44:2913-2933.

105. Reichheld, Frederick F. & Sasser, W. Earl Jr. (1990). ‘Zero Defections: Quality Comes to Services’. Harvard Business Review. September-October: 105-111.

106. Retzlaff-Roberts, D. (1996). ‘A ratio model for discriminant analysis using linear programming’. Eur. J. of Oper.. Res.94, 112-121.

107. Ritala, P. (2011). 'Coopetition Strategy--When is it Successful? Empirical Evidence on Innovation and Market Performance', British Journal of Management.

108. Ritala, P., K. Välimäki, K, Blomqvist, and K, Henttonen (2009). 'Intrafirm coopetition, knowledge creation and innovativeness', in Coopetition Strategy: Theory, Experiments and Cases, Taylor and Francis, 47,

109. Roth, K. and A. J. Morrison (1992). 'Implementing global strategy: characteristics of global subsidiary mandates', Journal of International Business Studies, pp. 715-735.

110. Sartori, G. (1970). 'Concept misformation in comparative politics', The American Political Science Review, 64, pp. 1033-1053.

111. Scherer, F. M. and D. Ross (2009). 'Industrial market structure and economic performance', ssrn.com.

112. Schertzer, C. B., & Schertzer, S. M. (2004). ‘Student satisfaction and retention: A conceptual model’. Journal of Marketing for Higher Education, 14(1), 79-91.

113. Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., Vespignani, A., & White, D. R. (2009). ‘Economic networks: The new challenges’, Science, pp. 422-425.

114. Seidman, Alan (2005). ‘Where we go from here: a retention formula for student success’. In Seidman, Alan, editor, College Student Retention: Formula for Student Success. 350 pg. US: Praeger Publishers.

115. Simmons, M. (2013). ‘If You Want To Go Fast, Go Alone. If You Want To Go Far, Go Together’, Forbes.

116. Soteriou, A. C., Karahanna, E., Papanastasiou, C., & Diakourakis, M. S. (1998). ‘Using DEA to evaluate the efficiency of secondary schools: the case of Cyprus'. International Journal of Educational Management, 12(2), 65-73.

117. Zahn, C. T. (1971). ‘Graph-theoretical methods for detecting and describing gestalt clusters’. Computers, IEEE Transactions on, 100(1), 68-86.

118. Teichmann, M. (2010). Personal interviews.

119. Tinto, Vincent (1986). ‘Theories of student departure revisited’. In Smart, J., editor, Higher education: A handbook of theory and research, Vol. 2: 359-384. Agathon.

120. Torim, A. and K. Lindroos (2008). 'Sorting Concepts by Priority Using the Theory of Monotone Systems.' in Conceptual Structures: Knowledge Visualization and Reasoning, Springer, pp. 175-188.

121. Tsai, W (2002). ‘Social Structure of "Coopetition" within a Multiunit Organization: Coordination, Competition, and Intraorganizational Knowledge Sharing’, Organizational Science, 13, pp. 179-190.

122. TUTMEKTORY (2014). MEKTORY of Talllinn University of Technology. 08.03.2014: http://www.ttu.ee/projects/mektory-eng/

123. TUTTechSchool (2014). Tech School of Tallinn University of Technology. 08.03.2014: http://www.ttu.ee/kooliopilasele/tehnoloogiakool/#

124. Ubi, J. (2003). 'General Manager's Role in Balancing Subsidiary Between Internal Competition and Knowledge Sharing', Unpublished thesis, University of Tartu, http://arxiv.org/submit/539275/preview.

125. Walker, D., J. Dauterive, E. Schultz and W. Block (2004). 'The Feminist Competition/Cooperation Dichotomy', Journal of Business Ethics, 55, pp. 243-254.

126. Walley, K. (2007). 'Coopetition: An introduction to the subject and an agenda for research', International Studies of Management and Organization, 37, pp. 11-31.

127. Wang, J., Yu, B., & Gasser, L. (2002). ‘Classification Visualization with Shaded Similarity Matrix’. In IEEE Visualization.

128. Varga, Liz (2009). ‘The coevolution of the firm and the supply network: a complex systems perspective’. Cranfield University: PhD. Thesis.

129. Wejnert, C. (2010). ‘Social network analysis with respondent-driven sampling data: A study of racial integration on campus’, Social Networks, 32, pp. 112-124.

130. Villanueva, Julian & Hanssens, Dominique M. (2006). ‘Customer equity: measurement, management and research opportunities’. 99 pg. now Publishers Inc.

131. Wishart, D. (1999). ‘Clustan Graphics3 Interactive Graphics for Cluster Analysis’. In ‘Classification in the Information Age’ (pp. 268-275). Springer Berlin Heidelberg.

132. VME, 2011. ‘Visual Matrix Explorer source code’. 08.03.2014: http://sourceforge.net/projects/vismatexplorer/

133. Vohandu L., Krusberg H. (1977) ‘A Direct Factor Analysis Method’, The Proceedings of TTU, 426, pp.11-21.

134. Vyhandu, L. (1980). ‘Some methods to order objects and variables in data systems’. Trans of Tallinn University of Technology, 482, 43-50.

135. Yami, S., Castaldo, S., Dagnino, B., & Le Roy, F. (Eds.). (2010). ‘Coopetition: winning strategies for the 21st century’. Edward Elgar Publishing.

136. Yan, L. & Fassino, M. & Baldasare, P. (2005). ‘Predicting customer behavior via calling links’. Proc. of Int. Joint Conf. on Neural Networks

137. Yankee gr. (2001). ‘Churn management in the mobile market’. PubID:YANL696399.

APPENDIX A

Übi, J.; Liiv, I.; Übi, E.; Võhandu, L. (2013). An analysis of community structure detection for educational coopetition. The 2nd IEEE International Conference on E-Learning and E-Technologies in Education (ICEEE2013), Lodz, Poland, September 23-25, 2013. IEEE, 2013, 104 – 109

An analysis of community structure detection for educational coopetition

Jaan Ubi
Department of Informatics
Tallinn University of Technology
Tallinn, Estonia
jaan.ubi@ttu.ee

Innar Liiv
Department of Informatics
Tallinn University of Technology
Tallinn, Estonia
innar.liiv@ttu.ee

Evald Ubi
Department of Finance and Economics
Tallinn University of Technology
Tallinn, Estonia
evald.ubi@ttu.ee

Leo Vohandu
Department of Informatics
Tallinn University of Technology
Tallinn, Estonia
leo.vohandu@ttu.ee

***Abstract*— The goal of the paper is to study how the strictly optimal solutions of community detection, based on similarity matrices, depend on the parameter of the distance threshold setting method, applied beforehand. In order to detect communities, we apply the oft-used modularity metric and arrive at strict optimality by linear programming, solving an np-hard problem. The distance threshold method is used, making the matrix more and more sparse, and thus the best value of the threshold is determined, by analyzing the number of subsequent clusters detected. Our method is applied on educational coopetition data in the business school of TUT, with four specializations, out of which we sample 36 students, each time selecting from a pair of specializations. Since the optimal number of clusters tends to be four, for any two-fold sampling, we detect a natural division within each specialization as well, the reason for which is a matter for further study. As a result, coopetition - the simultaneous competition and cooperation - is measured between the departments of the business school. The average grade of the students is a proxy for the competitive score of the department. The traditional conductance is used as a proxy for the cooperative score of the department. For our data, the optimal value for the threshold in community detection is 0.07, this way enough noise has been removed from the data, but not too many values, so that vital information is retained. Thus, we most often obtain our goal of detecting four clusters, in the two-fold sampling, effectively displaying the usefulness of fine-tuning the distance threshold while evaluating it by the strictly optimal community detection.*

***Keywords*— coopetition, metrics, social network analysis, linear programming, threshold method, education**

I. Introductory overview of the numerical method

Starting with early works such as [23] network science has proposed a number of community detection measures and algorithms [14][5]. Modularity measure was put forth by [13] and it has been proven to be highly effective in practice for community evaluation [6]. The strict optimization of modularity has been proven to be NP-hard [4] and highly effective heuristics have been devised e.g. [2]. This paper aims to explore the relationship between strictly optimal solutions and parameters that are used for methods refining the community detection data.

While visualizing the adjacency matrices of graphs, matrix reordering/seriation methods permute the rows of a community together, thus forming a strong block diagonal that visualizes the intracommunity ties with intercommunity ties being displayed off-diagonal (see [10] for a review and [9] for a software tool that implements a number of algorithms). A paper by [22] visualizes such a reordered shaded similarity matrix (by applying a method by [8]) and applies the distance threshold method, with different parameter values, by making the matrix more and more sparse (see Fig. 1). This way, it is hoped, that the irrelevant, immaterial noise is removed from the matrix, and only the information pertinent for community formation is retained. We aim to study such signal-to-noise ratio. The goal of the paper is - by raising the value of the distance threshold parameter and at each step solving the np-hard, strictly optimal modularity detection problem - to determine the optimal value for the threshold, as we perform supervised learning.

II. Details of modularity and distance threshold method

[13], in essence, defines modularity as

$$Q = \sum\left(A_{ij} - E_{ij}\right) * X_{ij} \rightarrow max \qquad (1)$$

ISBN: 978-1-4673-5094-5/13/$31.00 ©2013 IEEE

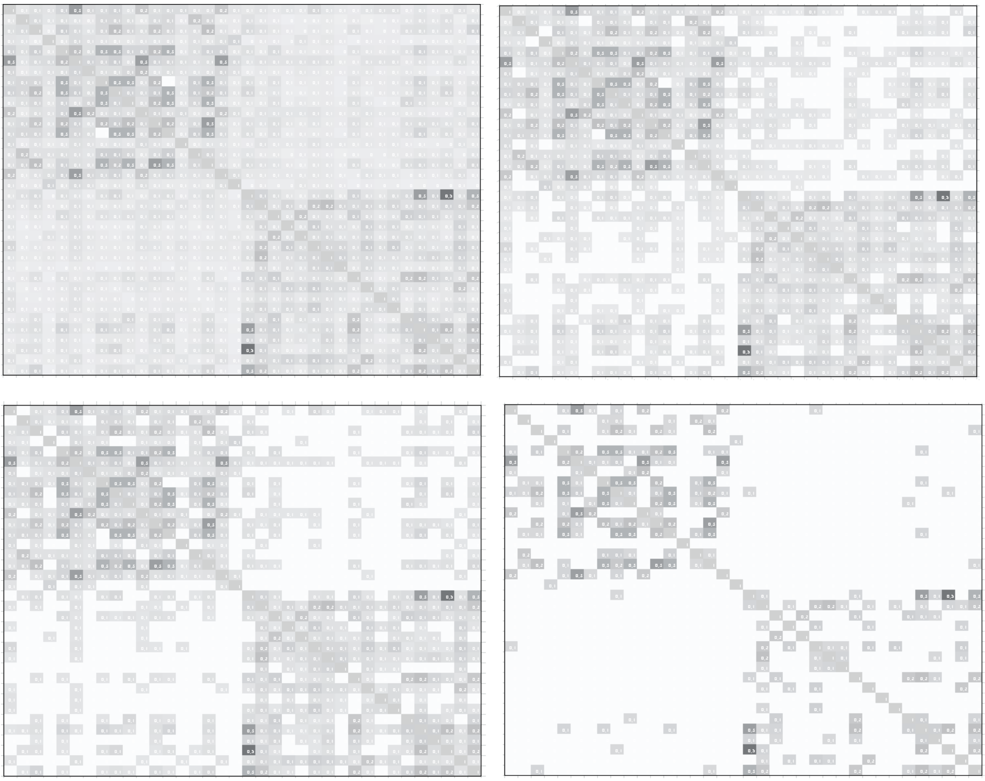


Fig. 1. The use of the distance threshold method – a successively more sparse adjacency matrix with its two communities permuted together. We study how this threshold influences the strictly optimal community detection.

where A is the adjacency matrix of a graph, Eij is the expected value of the element of the adjacency matrix and Xij is a binary variable, indicating whether the ith and the jth node of the graph are connected in the same community. The expected value, Eij, defined as

$$E_{ij} = \frac{\sum_{l=1}^{n} A_{il} * \sum_{k=1}^{n} A_{kj}}{\sum_{k=1,l=1}^{n} A_{kl}} \tag{2}$$

is row sum*column sum/matrix sum. This means that, basically, modularity has been carved after the Chi-Square characteristic, which also expects big values to appear in an element, that has big row and column sums, for example.

Communities are formed in such a way, that taking into account the whole matrix, nodes are connected, where the respective value in the adjacency matrix is bigger than its expected value.

The work by [4], determining the NP-hard nature of modularity, describes it as an integer linear programming (ILP) model. For the communities detected, it has three types of constraints: reflexivity, symmetry, transitivity,

$$X_{ii} = 1 \tag{3}$$

$$X_{ij} = X_{ji} \tag{4}$$

$$\begin{cases} X_{ij} + X_{ju} - 2 * X_{iu} \leq 1 \\ X_{iu} + X_{uj} - 2 * X_{ij} \leq 1 \\ X_{ju} + X_{ui} - 2 * X_{ji} \leq 1 \end{cases} \tag{5}$$

which ensure that entire communities are selected only. Specifically, the transitive nature of the problem sees to it, that if the first node is connected to the second, and that in turn to the third, the first has to also be connected to the third. Altogether, there are n^2 decision variables, n reflexivity constraints, n^2 symmetry constraints. The number of transitivity constraints is equal to the number of 3-combinations of n - but note that there are also three constraints that comprise each transitivity constraint. This way the total number of transitivity constraints is also equal to (the number of 3-permutations of n)/2 as the symmetric permutations can be pruned because of the symmetricity constraint. For example, a

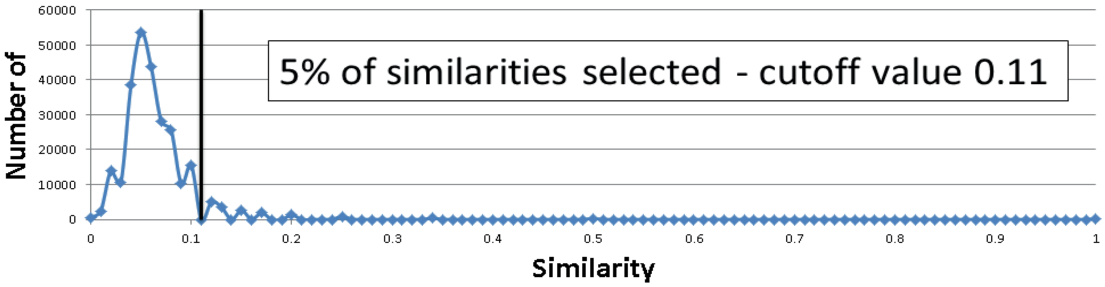


Fig. 2. Histogram of the student similarity matrix – similarity being defined as the reciprocal value of the Hamming distance of the student-by-course matrix

problem with 36 nodes, will have 1296 binary decision variables and 22752 (36+1296+21420) constraints.

The distance threshold method has been applied in [22]. The method essentially removes all distance values lower than a given threshold from the matrix. We raise this threshold gradually in subsequent iterations, and thus make the matrix more and more sparse, hoping to remove noise but not the signal from the data. In order to select the proper range for the distance threshold, the histogram of the data is constructed.

III. Educational Coopetition

The term coopetition - simultaneous competition and cooperation - first came into research focus in the second half of 1990s [3]. Coopetition can occur on many levels - societal, industrial cluster, market, organizational as well as personal - but it has so far predominantly been studied on the markets, e.g. [15][16]. This study is part of research effort to investigate coopetition on organizational level [17][21][19][20]. As we are investigating metrics that are applied for university management, parallels can be brought with huge multinational corporations (MNCs). Those operate with a divergent, partially overlapping structural map [20] on which structural solution World Product Mandate [18] (being mandated) is an embodiment of internal competition that takes place and solution Center-of-Excellence [12] (becoming a center) an embodiment of the simultaneous cooperation. In literature coopetition is implied even if either competition or cooperation seems to prevail - thus for instance cases such as those propounding "coevolving systems" (in favor of cooperation) [7] or "internal markets" (in favor of competition) [1], really also have some flavor of the other side of the duality mixed in.

As for the universities, departments of, say, the business school, have mandates to teach their students - mandates showing the competitive position, as is evident in the average grade of the students. At the same time they can be conceptualized as centers as the curriculum is intertwined - this being evident in the way students of one specialization take courses from another department. This way departmental capabilities are developed, which bring with it the unit becoming better at both its primary goal (grooming its own students) as well as the secondary goal (having a curricula, which is useful for other parts of the school) - the two together maximizing the performance of the entire organization [20].

IV. Data on Coopetition

The data on coopetition comes from Tallinn School of Economics and Business Administration (TSEBA) of Tallinn University of Technology (TUT). It consists of 331 students, who have graduated the curriculum "Business" during the time span 1997-2010. We have information regarding the courses that the students have attended, the grades they received, as well as on the departmental final specialization chosen. Altogether, students have attended 759 courses. The course selection data follows the power law. The four specializations of TSEBA are: finance, marketing, accounting and management (see Table 1).

TABLE I. THE STUDENTS AND THE SPECIALIZATIONS OF THE BUSINESS SCHOOL TSEBA STUDIED.

Specialization	Number of students
Marketing	87
Finance	74
Management	71
Accounting	99
Total	**331**

As the first step of analysis, a binary matrix - with rows representing the students and columns representing the courses attended - is constructed.

This matrix is about 95% sparse. Next, a 331x331 distance matrix between all the students is formed. At first, Hamming distances are used - counting the total number of different courses selected by the two students. The second step involves the calculation of the similarity matrix, by taking the reciprocal value of the distance. In order to later select the threshold, we also calculate the histogram of the similarities (see Fig. 2).

V. An Application of the Method

We begin by applying a fast modularity heuristic [2] to the entire 331x331 similarity matrix - in its original form (threshold 0), as well as with 95% of smallest values removed (threshold 0.11, cf. the histogram in Fig. 2). While the full similarity matrix has the departmental specialization of students detected badly, resulting in 16 mixed communities detected; sparse matrix ends up having 8 communities, students of each specialization being divided between two communities and all communities begin comprised of only one type of students.

As heuristic modularity algorithms have potential community detection problems of intrinsic nature, we want to abstract ourselves from such and hereon look at the detection of communities backed by the strict optimality of the modularity criteria.

As we use Gurobi engine of Analytic Solver Platform and require to solve well over thousand problems, our graph size chosen has 36 nodes; thus being with 1296 binary decision variables and 22752 (36+1296+21420) constraints. This problem has a solution time of roughly 90 seconds.

As there are only 36 nodes in the graph, we sample 18 each from two specializations, for one optimization. There are six combinations to sample two out of four specializations. We sample each specialization combination 8 times. We vary the threshold criteria from 0 to 0.12 by 0.01 steps. Thus at this step we perform 6*8*13=624 optimizations.

The results are depicted on Table 2, showing the average number of pure clusters obtained. In about 5% of optimizations the result had one cluster that was not "pure" - that is it contained more than one type of specializations - these optimizations are omitted from the results. For different threshold levels, the average number of clusters varies between 4.56 and 10.90, being at its worst levels when the threshold is either too low or too high. Clearly the optimal threshold level, for our data, is 0.07.

Next, we look at the fact, that there are four clusters in each optimization, not two - while at the same time, the clusters are almost always "pure". We study the subcluster stability within each specialization. In order to do that we set the threshold level at 0.07. We use all the six combinations of specializations. We sample each combination 100 times. This way we now perform 1*6*100=600 optimizations.

The results are depicted on Figure 3, showing very good stability. For all four specializations, the two internal subclusters (which can be seen from the "checkerboard" pattern) can easily be permuted together, leaving very little noise off-diagonal. The reason for the emergence of such internal subclusters, however, is a topic for a follow-up research. It can tentatively be surmised, though, that perhaps the reason is the revamp of the curriculum, undertaken to move on with the Bologna Process.

VI. Coopetition score of the departments

Having detected the departmental specializations, we next measure their coopetitive stance, as we have done in [20], by using the current data table.

The competitive position of a department is evident in the average grade of its students. For the cooperative position we look at traditional conductance of social network analysis [11]:

$$Cooperation = \frac{\sum linkages\ from\ the\ specialization\ leading\ to\ others}{\sum all\ intragroup\ linkages\ for\ the\ specialization} \quad (6)$$

This uses students as proxies, for showing how interlinked the curriculum is. If there are enough Management students taking Finance courses, it will be evident in the similarity matrix. We normalize both scores as percentages of the maximum, and calculate the coopetition score as an equally weighted average of the two.

TABLE II. The average number of "pure" clusters obtained, dependent on the threshold selected, across all the two-fold sampling combinations of the four specializations; 18+18 students were selected for each optimization

Threshold	Specialization pair						
	Marketing-Finance	Marketing-Management	Marketing-Accounting	Finance-Management	Finance-Accounting	Management-Accounting	Grand Total
0	7,88	8,75	7,5	8,88	9,13	9,75	**8,65**
0,01	7,88	8,75	7,5	8,88	9,13	9,75	**8,65**
0,02	7,88	8,75	7,5	8,88	9,13	9,75	**8,65**
0,03	7,88	8,75	7,5	8,88	9,13	9,75	**8,65**
0,04	7,88	8,63	7,5	8,88	9	9,63	**8,58**
0,05	6	6,63	6	7	7,88	7,75	**6,88**
0,06	4,38	4,13	4,75	5,13	6	5,25	**4,94**
0,07	**4,13**	**4,25**	**4,5**	**4,63**	**5,13**	**4,75**	4,56
0,08	4,63	5,25	5	5,5	5,88	5,38	**5,27**
0,09	5,13	5,75	5,75	6,38	6,75	6,13	**5,98**
0,1	6,13	6,88	6,63	7,63	7,13	7,38	**6,96**
0,11	7,88	8,75	7,88	9,25	8,75	9,38	**8,65**
0,12	10,5	11,13	9,75	11,38	10,13	12,5	**10,9**

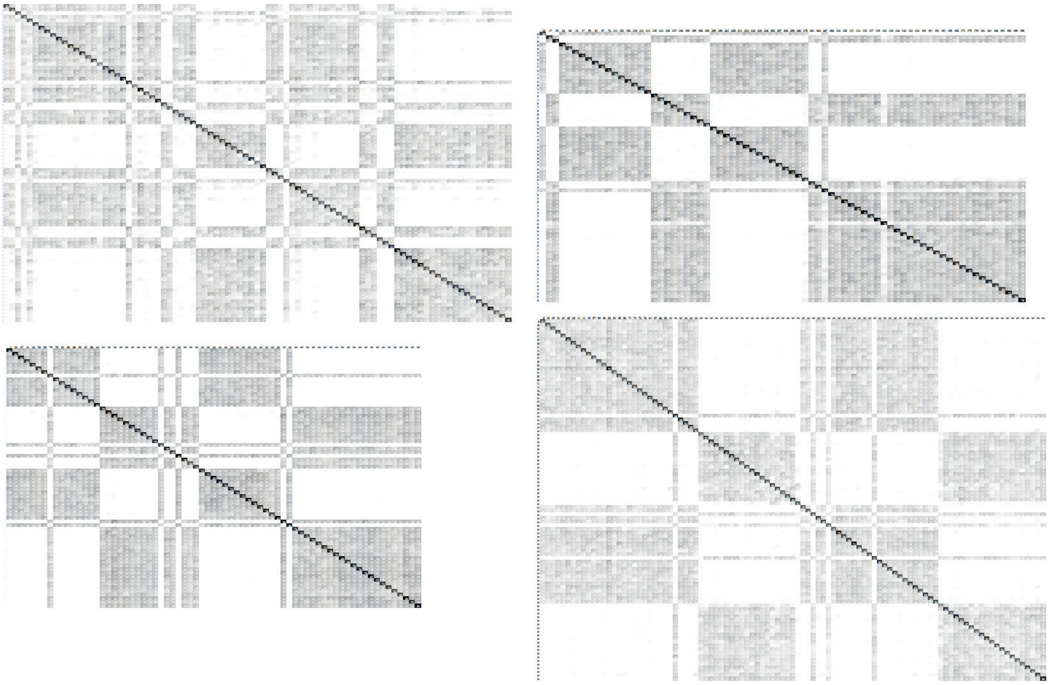


Fig. 3. The subcluster stability for the four specializations – Marketing, Finance, Management and Accounting, each matrix showing the "checkerboard pattern", which can be permuted together, containing two clusters and very little off-diagonal noise.

The results are summarized in Table 3. The competitive position is the strongest for the Finance specialization. The curriculum of Management (which does worst in competitive terms), is the most interlinked with the others. The equal weighted score is the highest for Finance at 98%, but Management comes in as close second at 93%. Thus, our goal as a university would be to financially reward the coopetitive outcome, as we can ground our footing having detected the communities in a more sound manner.

TABLE III. THE COMPUTATION OF THE COOPETITION SCORE.

Specialization	Average grade (**competititiveness**)	Conductance (**cooperativeness**)	Average grade/ maximum average grade	Conductance/ maximum conductance	(EQUAL-WEIGHTED RELATIVE) **CO-OPETITION SCORE** (Average grade/maximum average grade+ Conductance/maximum conductance)/2
MARKETING	**3,38**	**12,78%**	92,86%	37,69%	**65,27%**
MANAGEMENT	**3,15**	**33,91%**	86,54%	100,00%	**93,27%**
FINANCE	**3,64**	**32,85%**	100,00%	96,87%	**98,44%**
ACCOUNTING	**3,35**	**15,75%**	92,03%	46,45%	**69,24%**
maximum	3,64	33,91%			

VII. CONCLUSION

This paper has combined the strict optimization of the modularity metric with the use of distance threshold method, showing the importance of discarding the noise, but retaining the signal. As the problem is NP-hard, data has been repeatedly sampled from the whole set. The communities detected this way have proven to be highly stable and more than 99% of all the clusters detected have been "pure" in the sense of being composed of only one specialization. We have found the optimal threshold level for our data to be 0.07. At the same time the detection of clusters has been performed in a fractal-like self-similar fashion, discovering them on a small scale and finding validation in the whole data set. The detection of communities has served the purpose of calculating the coopetition score for a department. This has shown to be dependant upon competitive well-being, as in the case of Finance, but also strongly influenced by how well the department serves the rest of the business school, as in the case of Management.

REFERENCES

[1] Birkinshaw, Julian. Entrepreneurship in the global firm. Thousand Oaks: Sage, 2001.

[2] Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. "Fast unfolding of communities in large networks." Journal of Statistical Mechanics: Theory and Experiment 2008, no. 10 (2008): P10008.

[3] Brandenburger, A. M., and B. J. Nalebuff. "Co-opetition." (1996).

[4] Brandes, Ulrik, Daniel Delling, Marco Gaertler, Robert Gorke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. "On modularity clustering." Knowledge and Data Engineering, IEEE Transactions on 20, no. 2 (2008): 172-188.

[5] Chen, Jiyang, Osmar R. Zaïane, and Randy Goebel. "Detecting Communities in Social Networks Using Max-Min Modularity." In SDM, vol. 3, no. 1, pp. 20-24. 2009.

[6] Danon, Leon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. "Comparing community structure identification." Journal of Statistical Mechanics: Theory and Experiment 2005, no. 09 (2005): P09008.

[7] Eisenhardt, Kathleen M., and D. Charles Galunic. "Coevolving: At last, a way to make synergies work." Harvard Business Review 78, no. 1 (2000): 91-102.

[8] Gale, Nathan, William C. Halperin, and C. Michael Costanzo. "Unclassed matrix shading and optimal ordering in hierarchical cluster analysis." Journal of Classification 1, no. 1 (1984): 75-92.

[9] Liiv, Innar, Rain Opik, Jaan Ubi, and John Stasko. "Visual matrix explorer for collaborative seriation." Wiley Interdisciplinary Reviews: Computational Statistics 4, no. 1 (2012): 85-97.

[10] Liiv, Innar. "Seriation and matrix reordering methods: An historical overview." Statistical analysis and data mining 3, no. 2 (2010): 70-91.

[11] Mislove, Alan, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. "Measurement and analysis of online social networks." In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, pp. 29-42. ACM, 2007.

[12] Moore, Karl, and Julian Birkinshaw. "Managing knowledge in global service firms: centers of excellence." The Academy of Management Executive 12, no. 4 (1998): 81-92.

[13] Newman, Mark EJ, and Michelle Girvan. "Mixing patterns and community structure in networks." In Statistical Mechanics of Complex Networks, pp. 66-87. Springer Berlin Heidelberg, 2003.

[14] Newman, Mark EJ. "Analysis of weighted networks." Physical Review E 70, no. 5 (2004): 056131.

[15] Peng, Tzu ‐ Ju Ann, and Mike Bourne. "The Coexistence of Competition and Cooperation between Networks: Implications from Two Taiwanese Healthcare Networks*." British Journal of Management 20, no. 3 (2009): 377-400.

[16] Ritala, Paavo. "Coopetition strategy–When is it successful? Empirical evidence on innovation and market performance." British Journal of Management 23, no. 3 (2012): 307-324.

[17] Ritala, Paavo, Kari Välimäki, Kirsimarja Blomqvist, and Kaisa Henttonen. "Intrafirm coopetition, knowledge creation and innovativeness." Co-opetition Strategy—Theory, Experiments and Cases (2009): 64-73.

[18] Roth, Kendall, and Allen J. Morrison. "Implementing global strategy: characteristics of global subsidiary mandates." Journal of International Business Studies (1992): 715-735.

[19] Tsai, Wenpin. "Social structure of "coopetition" within a multiunit organization: Coordination, competition, and intraorganizational knowledge sharing." Organization science 13, no. 2 (2002): 179-190.

[20] Ubi, Jaan, Liiv, Innar, Varblane, Urmas and Vohandu, Leo. "Measuring Multinational Corporation-like internal co-opetition duality in a university context". (2013): http://arxiv.org/abs/1208.5438.(in press).

[21] Walley, Keith. "Coopetition: An introduction to the subject and an agenda for research." International Studies of Management and Organization 37, no. 2 (2007): 11-31.

[22] Wang, Jun, Bei Yu, and Les Gasser. "Classification Visualization with Shaded Similarity Matrix." In IEEE Visualization. 2002.

[23] Zahn, Charles T. "Graph-theoretical methods for detecting and describing gestalt clusters." Computers, IEEE Transactions on 100, no. 1 (1971): 68-86.

APPENDIX B

Übi, J.; Übi, E.; Liiv, I.; Võhandu, L. (2013). Predicting student retention by comparing histograms of bootstrapping for Charnes-Cooper transformation-linear programming discriminant analysis. The 2nd IEEE International Conference on E-Learning and E-Technologies in Education (ICEEE2013), Lodz, Poland, September 23-25, 2013. IEEE, 2013, 110 – 114

Predicting student retention by comparing histograms of bootstrapping for Charnes-Cooper transformation-linear programming discriminant analysis

Jaan Ubi
Department of Informatics
Tallinn University of Technology
Tallinn, Estonia
jaan.ubi@ttu.ee

Evald Ubi
Department of Finance and Economics
Tallinn University of Technology
Tallinn, Estonia
evald.ubi@ttu.ee

Innar Liiv
Department of Informatics
Tallinn University of Technology
Tallinn, Estonia
innar.liiv@ttu.ee

Leo Vohandu
Department of Informatics
Tallinn University of Technology
Tallinn, Estonia
leo.vohandu@ttu.ee

***Abstract*— The goal of the paper is to predict student retention by using linear discriminant analysis with bootstrapping. The result (93%) provides accuracy superior to the bootstrapping of a comparative method, as well as to the non-bootstrapping variations. In order to perform discriminant analysis, we linearize a fractional programming method by using Charnes-Cooper transformation and apply linear programming, while the comparative approach uses deviation variables to tackle a similar multiple criteria optimization problem. We train the discriminatory hyperplane family and apply it to the testing set – thus arriving at a set of histograms. We analyze the histograms by using the simple mean – best for prediction – and a five-fold Kolmogorov-Smirnov test – best used for resources allocation, in order to act on the final results. Final results are the outcome of applying the hyperplane family on freshman data.**

Keywords*— Discriminant analysis; linear programming; Charnes-Cooper transformation; Data Envelopment Analysis; bootstrapping; Kolmogorov-Smirnov test; histogram; data mining; student retention; student dropout; churn;

I. Introduction

Classification methods used in data mining include discriminant analysis, support vector machines, decision trees, neural networks, gene expression programming and others. As linear methods are generally superior performance wise, we seek to apply the simplex method of linear programming, for performing linear discriminant analysis – this instead of oft-chosen Fisher's linear discriminant.

A number of approaches have been proposed for this, including [1], [2], [3], [4], [5], [6], [7]. We follow the call for applying variations of Data Envelopment Analysis (DEA), by [7]; and compare it to a more advanced deviation variable approach [6].

We apply data mining, in order to predict student retention. As the data preprocessed and used for training our model determines the result by using only 14 variables and 400 cases, we can furthermore apply bootstrapping and create a family of discriminatory hyperplanes. Each case will then correspondingly have a histogram of its estimations, on which we will base our decision.

For comparing the final results, we perform a five-fold Kolmogorov-Smirnov test on the histograms, look at the simple average of the histogram and also at the results of a single iteration of bootstrapping. Besides training and testing, we also apply the model on a set of freshman data.

II. MULTIPLE CRITERIA LINEAR PROGRAMMING (MCLP) APPOACH FOR TWO CLASS DISCRIMINANT ANALYSIS

A. Problem description

Linear discriminant analysis is used in order to find a hyperplane that separates the two sets of students, in the best way achievable, thus we have:

$$A_iX \le b_i, A_i \in G_1$$

$$A_iX \ge b_i, A_i \in G_2$$

In order to do that, we find the correct and the erroneous distance that each data point has to the hyperplane (see Figure 1). Correct distances are denoted by β and erroneous distances by α, which we add to each equation. The signs of β and α, will depend on whether the student will drop out, or graduate. Thus our objective is to simultaneously maximize the sum of β and to minimize the sum of α and we arrive at the following set of equations (note that for one student, only either α or β will be different from zero, depending on the accuracy of the

ISBN: 978-1-4673-5094-5/13/$31.00 ©2013 IEEE

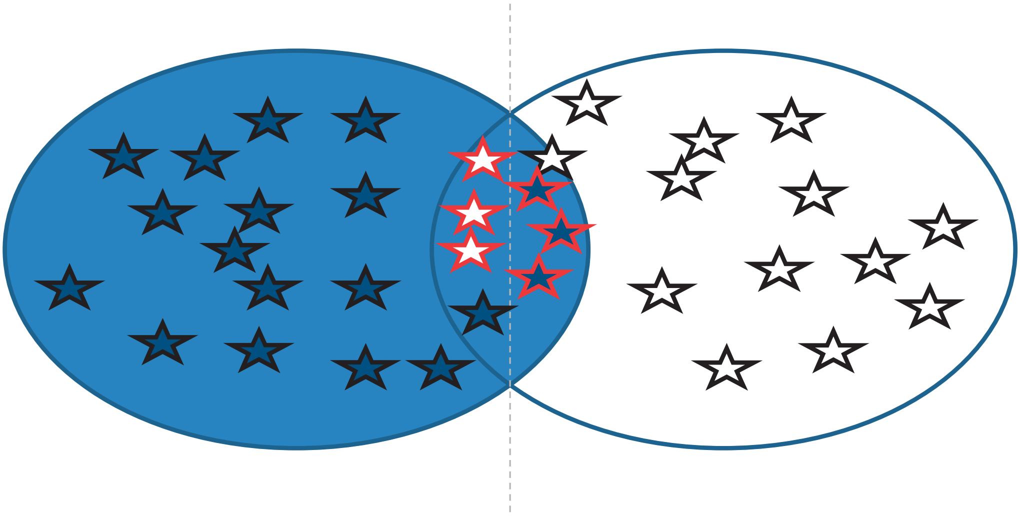


Figure 1. Separating hyperplane, the correct and erroneous distances of data points.

prediction – or both will be zero, if the student lies on the discriminatory hyperplane):

$$Z_1 = \sum \beta_i \rightarrow max$$

$$Z_2 = \sum \alpha_i \rightarrow min$$

$$A_i X = b_i + \alpha_i - \beta_i, \forall A_i \in G_1$$

$$A_i X = b_i - \alpha_i + \beta_i, \forall A_i \in G_2$$

$$\alpha_i, \beta_i \geq 0$$

B. Charnes-Cooper transformation approach to the MCLP

In order to solve the problem with simplex method, we need to have a single objective function. We use the following fractional programming problem, where the numerator is essentially a vector consisting of all β-s and α-s, with the coefficients of α-s being zero; and the denominator is a vector consisting of all β-s and α-s, with the coefficients of β-s being zero. We once again add β-s and α-s to inequalities of the constraints, with corresponding sign changes (not explicitly indicated in the following equations).

$$z = \frac{cx + c_0}{dx + d_0} \rightarrow max$$

$$Ax \leq b$$

$$x \geq 0$$

Close variations of this problem were discussed in [7] and the article ended with a call to deploy Data Envelopment Analysis. Thus we linearize the problem by applying Charnes-Cooper transformation, the use of which has originated with DEA, in the following way. We denoto y0 as:

$$y_0 = \frac{1}{dx + d_0}$$

and xj as:

$$x_j = \frac{y_j}{y_0}$$

Thereby arriving at the following single criteria linear programming problem:

$$z = cy + c_0 y_0 \rightarrow max$$

$$Ay \leq b y_0$$

$$dy + d_0 y_0 = 1$$

$$y \geq 0$$

This has decision variables that have been transformed, and an additional constraint that equals to 1. Traditionally we would need a constant d0, that avoids the possibility of dividing by zero, but not in this case, as there is virtually no chance of our data being discriminated perfectly. We will further on use variable y0 for configuring the Charnes-Cooper transformation LDA model.

III. A COMPARATIVE APPROACH TO THE MCLP DISCRIMINATION – THE USE OF DEVIATION VARIABLES

Paper [6] puts forth a comparative approach to data mining MCLP discriminant analysis problems. It uses deviation variables, which take the absolute value of the left side of the equation. As sums of α-s are added to, and sums of β-s are subtracted from a constant we can now include both in a single objective function, as follows:

$$Z = d_\propto^- + d_\propto^+ + d_\beta^- + d_\beta^- \rightarrow min$$

$$\propto_* + \sum \propto_i = d_\propto^- + d_\propto^+$$

$$\beta_* - \sum \beta_i = d_\beta^- + d_\beta^+$$

ISBN: 978-1-4673-5094-5/13/$31.00 ©2013 IEEE

$$A_iX = b_i + \alpha_i - \beta_i, \forall A_i \in G_1$$

$$A_iX = b_i - \alpha_i + \beta_i, \forall A_i \in G_2$$

$$\alpha_i, \beta_i \geq 0$$

The parameters to be configured in this model ara β* and α*.

IV. STUDENT RETENTION DATA FROM TALLINN UNIVERSITY OF TECHNOLOGY

Our database consists of course declaration data of Tallinn University of Technology (TUT). In the time span of 1997-2010, 1.3 million course attendances have been registered – in total by 40 000 students. We focus on the graduates of Tallinn School of Economics and Business Administration (TSEBA) of TUT. We are looking at students, who have finished their freshman year of studies – and have thus attended the same 10 courses. There are all in all 928 such students, 633 of whom will be selected by us – the criteria being that the student started his studies at least four years ago, so that there would have been sufficient time for graduation. We will finally apply our models on the remaining 295, in order to predict, whether they are going to graduate. The 633 students in turn were divided into 425, who successfully graduated, and 208, who did not.

In addition to the attendance data, we know whether the student is a he or a she, what is the persons age at the time of accession to the university, what is the mother tongue, whether it is a full-time or a part-time student, and how good is the previous educational institution – the high school.

The mother tongue is predominantly either Estonian or Russian, correspondingly to our demographic division.

Every year a leading weekly, Eesti Ekspress, publishes high school ratings, showing the average performance of recent high school graduates in the national graduation examinations.

All the data was normalized and discretized to integer values 0..5, while preparing for the discriminant analysis.

While analyzing the distributions of the 10 courses taken, it turned out that one of the courses resulted in all the students passing. This course was removed from the data, as it would have resulted in unsuccessful optimizations. Furthermore, one course had only 14 failed results. The probability that we would draw 400 times out of 633, without ever hitting any of the 14 failed ones was calculated to be virtually nonexistent, and thus this course was not removed (see the exact nature of optimizations below, for an explanation). Nevertheless such situation occurred during one optimization, the result of which was removed.

4. ANALYZING THE HISTOGRAMS OF BOOTSTRAPPING THE OPTIMIZATION

A. The procedure

In order to perform bootstrapping about 2/3 of 633 students, about whom we knew the final outcome (namely, 400), were selected as training data. Our model of Charnes-Cooper transformation was optimized. As the next step, the remaining 233 students were entered into the model – as a testing sample – and the results calculated.

Bootstrapping was performed 65 times, and the results – both for training and testing – were recorded. This resulted in histograms (see Figure 2) – one for each student, calculated based on the 65 distances from the hyperplane. These histograms reflect the behavior of the student, as it was relating to the discriminatory hyperplane family that had been created.

In order to make the final assessment, whether the student would graduate, histograms were evaluated. At first a fivefold Kolmogorov-Smirnov test was undertaken. Average histograms of dropouts and graduates were calculated (see Figure 3). Kolmogorov-Smirnov test measures the maximum distance between cumulative distribution functions and tests

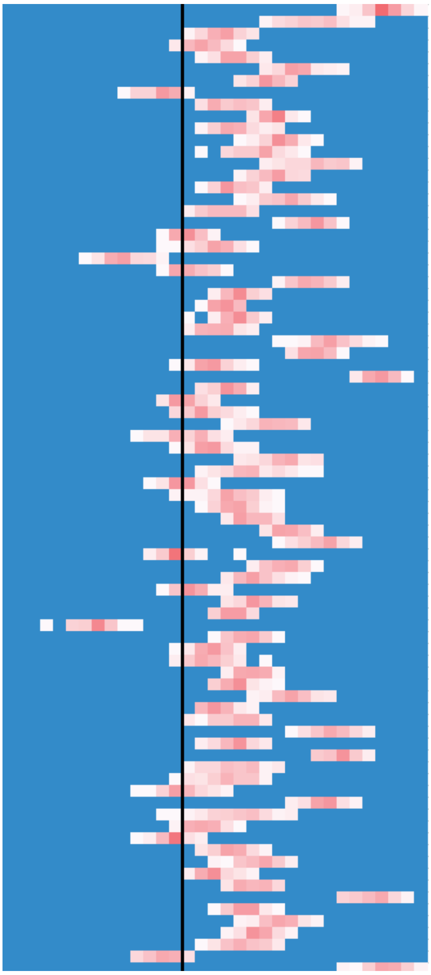


Fig. 2. Histograms resulting from bootstraping – some graduates.

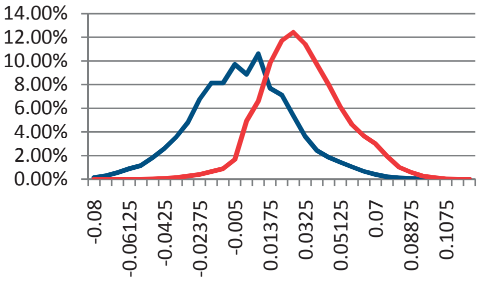


Fig. 3. Average histograms of dropouts and graduates

whether they have sufficient similarity. The maximum distance allowed depends on the sample sizes and the level of probability. In our case the level chosen – 95% – resulted in the maximum allowed difference of 24% (see Figure 4, for an example).

The following five KS tests were undertaken. Whether the student is sufficiently similar to the "profile" of the (average) graduate; whether he is sufficiently similar to the "profile" of the (average) dropout; whether his distribution function is to the "right" of the one of the graduate (very certain to graduate); whether his distribution function is to the "left" of the one of the dropout (very certain to drop out); or whether he is "between" the graduate and dropout profiles – "uncertain".

While analyzing the results (see Table 1), it was observed, that the test proved to be too "strong" – too many students were detected as "uncertain" or borderline – with only 60% of students predicted correctly. The test performed even worse than a random draw (about 66% of our students had graduated), and also worse than a single execution of the model (about 77% accuracy). However, it would be useful to be applied on data later on – as employees of the dean's office have to decide, on whose case to intervene. As the resources are limited, they might not be spent on those, who are most certain to drop out.

The final assessment, about the student graduating or dropping out, was made using the simple average of its histogram. If on the average, the student remained on the graduating side of the hyperplane family – he was predicted to finish his studies. This resulted in the overall accuracy of 93% (95% for the graduates and 90% for the dropouts).

Our model was compared to the histogram-bootstrap version of the deviation variable model – and performed in a superior fashion. In that model the overall accuracy achieved for the simple mean was 85%, thus proving the validity of our approach.

TABLE I. OVERALL ACCURACY OF THE MODELS

KS test	Random draw	Single iteration	Deviation variable	CC transformation
60%	66%	77%	85%	93%

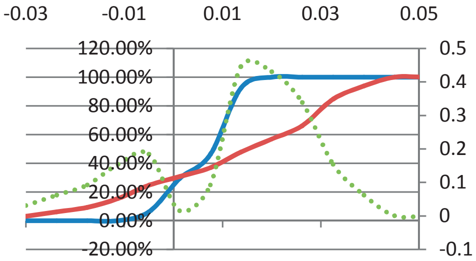


Fig. 4. Calculating the Kolmogorov-Smirnov test, with our sample sizes 425*65=27625 and 65 and at level 95%, critical difference allowed is 24%.

During the development we also tested our models by changing their free parameters. For example, we changed the y0 parameter of the Charnes-Cooper model, but without much success, except at some values, which were definitely out of range. Thus 0.01 was selected as the most obvious candidate. For deviation variable model α* was selected as 0.5 and β* as 30 000, in the same way.

Furthermore, we varied the ratio of graduates/dropouts in our training sample. However, the choice according to which the proportion of graduates was the same for the training and the total sample, worked by far in the most effective way. We also tried to, for example, train on 50%-50% graduation data, but that would have meant that we take the dropout information into account to a too extensive degree – and the model performed dismally.

While data preprocessing was performed in R; Excel, its VBA and Frontline Solver Platform was used for all other tasks. One bootstrapping execution took about 20 minutes.

Finally, the family of discriminatory hyperplanes was applied to the 295 students, who had only recently started their studies – for whom we needed the prediction. 70% of students were predicted to successfully finish their studies.

V. CONCLUSION

We have applied Charnes-Cooper transformation to a fractional programming problem, thus linearizing it. Furthermore, bootstrapping has been applied, resulting in a discriminatory hyperplane family – and a set of histograms for each student. Comparison is made with a deviation variable LDA method.

The data was preprocessed, normalized and discretized, the unit column discarded, statistical test undertaken for determining the probability of appearance of another unit column.

The final prediction is made, based on the simple average of the histogram. A five-fold Kolmogorov-Smirnov test produces results that can be used for resource allocation, while contacting students. The model was applied on freshman students and 70% of students were predicted to graduate.

The models gave the best results after some configuration, parameters such as y0, α* and β* had to be set and the proportion of graduates in the training set established. The best results emerged from our model with bootstrapping (93%), followed by the deviation variable model with bootstrapping (85%), the models in a single iteration (77%), the random choice (66%) and the models with bootstrapping using Kolmogorov-Smirnov test as the criteria (60%).

REFERENCES

[1] Bal, Hasan, H. Hasan Örkcü, and Salih Çelebioğlu. "An experimental comparison of the new goal programming and the linear programming approaches in the two-group discriminant problems." *Computers & Industrial Engineering* 50, no. 3 (2006): 296-311.

[2] Erenguc, S. Selcuk, and Gary J. Koehler. "Survey of mathematical programming models and experimental results for linear discriminant analysis." Managerial and Decision Economics 11, no. 4 (1990): 215-225.

[3] Freed, Ned, and Fred Glover. "Simple but powerful goal programming models for discriminant problems." European Journal of Operational Research 7, no. 1 (1981): 44-60.

[4] Glover, Fred. "Improved Linear Programming Models for Discriminant Analysis." Decision Sciences 21, no. 4 (1990): 771-785.

[5] Koehler, Gary J. "Considerations for mathematical programming models in discriminant analysis." Managerial and Decision Economics 11, no. 4 (1990): 227-234.

[6] Kou, Gang, Xiantao Liu, Yi Peng, Yong Shi, Morgan Wise, and Weixuan Xu. "Multiple criteria linear programming approach to data mining: Models, algorithm designs and software development." Optimization Methods and Software 18, no. 4 (2003): 453-473.

[7] Retzlaff-Roberts, Donna L. "A ratio model for discriminant analysis using linear programming." European journal of operational research 94, no. 1 (1996): 112-121.

APPENDIX C

Liiv, I.; Öpik, R.; Übi, J.; Stasko, J. (2012). Visual matrix explorer for collaborative seriation. Wiley Interdisciplinary Reviews: Computational Statistics, 4(1), 85 – 97

Advanced Review

Visual matrix explorer for collaborative seriation

Innar Liiv,[1]* Rain Opik,[1] Jaan Ubi[2] and John Stasko[3]

In this article, we present a web-based open source tool to support cross-disciplinary collaborative seriation with the following goals: to compare different matrix permutations, to discover patterns from the data, annotate it, and accumulate knowledge. Seriation is an unsupervised data mining technique that reorders objects into a sequence along a one-dimensional continuum to make sense of the whole series. Clustering assigns objects to groups, whereas seriation assigns objects to a position within a sequence. Seriation has been applied to a variety of disciplines including archaeology and anthropology; cartography, graphics, and information visualization; sociology and sociometry; psychology and psychometrics; ecology; biology and bioinformatics; cellular manufacturing; and operations research. Interestingly, across those different disciplines, there are several commonly emerging similar structural patterns. Visual Matrix Explorer allows users to explore and link those patterns, share an online workplace and instantly transmit changes in the system to other users. © 2011 John Wiley & Sons, Inc. *WIREs Comp Stat* 2011 DOI: 10.1002/wics.193

Keywords: seriation; matrix reordering; collaboration; information visualization

INTRODUCTION

This article introduces a web-based tool for exploratory visual analytics—Visual Matrix Explorer (VME)—that enables dynamic evaluation and visualization of matrices, and the linking of different seriation results. The main motivation for this article was to change the tradition of literature review being a static textual result and to allow for an interaction between different theories and methods presented in overviews. Current article is complementing a recent literature review[1] on seriation and matrix reordering. We presently use the tool to evaluate, visualize, and link different seriation results from different disciplines, but the real value is much broader due to seriation being one of the fundamental learning components[2] to organize events, objects or other phenomena we are looking to understand.

VME is a tool for researchers investigating surveys and questionnaires, social networks, process execution logs or other data that can be expressed in a tabular form. It supports dynamic theory building and comparison, allowing the user to interactively explore and link any rankings of importance, interestingness, and focus (from any theory)—and finally settle for a suitable interpretation. The dynamic nature of the tool is additionally manifested in the capability of focusing on a subset of data and running operations therein.

This article is not just about comparing the results and theories of seriation and clustering, but—using the terminology of information visualization and interaction—is concerned with cross-disciplinary and cross-theory brushing.[3] VME gives the user a possibility to inspect, whether a collection of facts—a meaningful knowledge in one discipline—can yield relevant structures in other theories, adds new opportunities for exploratory data analysis and knowledge discovery in general.

The article is organized in the following way. In the next section, the difference and interplay between seriation and matrix reordering is described, followed by a brief discussion. In Section '*Comparing Theories*', the objective of finding related traits in different theories is elaborated. In Section '*The Functionality of VME*', the functionality of the tool is described. Section '*Illustrative Examples from Different Disciplines*' contains several accounts of exploration from different disciplines, each with unique investigative and analytical tasks, but a shared general goal.

*Correspondence to: innar.liiv@ttu.ee

[1]Department of Informatics, Tallinn University of Technology, Tallinn, Estonia

[2]Computational Systems Biology Laboratory, University of Georgia, Athens, GA, USA

[3]School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA

DOI: 10.1002/wics.193

SERIATION AND MATRIX REORDERING

As stated in '*Introduction*' section, seriation assigns objects to positions in a sequence, according to some predefined principle and objective. An intuitive example would be a shopping list that some people write down before a visit to a store. This list can be compiled in multiple ways. Depending on the organization, it can serve different goals—either to contain just the original information (what to buy?) or to include a categorization and ordering (seriation) in some way more reasonable for shopping (e.g., to prioritize or to minimize the walking distance in a shop). An interested reader is referred to Belknap's recent book about ordering lists.[4] Seriating a data table is a natural two-dimensional extension to organizing lists, where both rows and columns can be reordered.

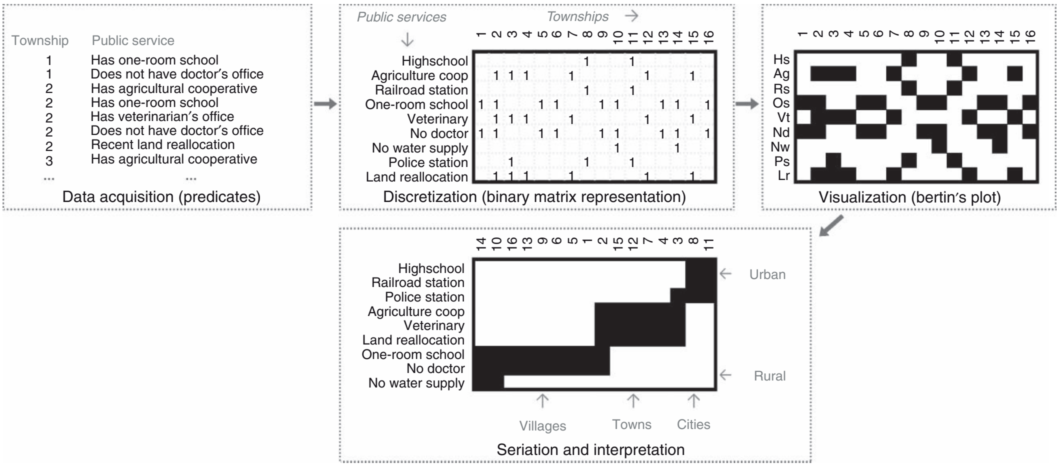
While new methods of manipulating and storing data are being developed and brought into mainstream, the table (matrix) with its method—the spreadsheet—remains the predominant form of storing data,[5] already, since the ancient times.[6]

When finding a seriation for a matrix we can, depending on the optimization objective, perform it independently for the rows and the columns as suggested by Lenstra,[7] or do it in a dependent manner recommended by Niermann.[8] The combinatorial optimization problem is essentially one of finding a permutation, thus being in a factorial search space and using (often greedy) heuristics. As Figure 1 depicts the process of seriation by the classical example of Bertin's townships,[9] it can be observed that the result makes the relationships and patterns within the dataset more evident—most importantly without any dimensionality or other reduction in the data. Note, that after the initial data has been discretized, visualization transforms ones into black dots with zeroes forming the white background, using coding similar to a seriation package in R environment.[10]

The research of seriation and matrix reordering methods go back more than 100 years.[1] However, with a few exceptions, the research has only concentrated on choosing the best static representation, whereas according to the information visualization studies, the interaction with the representation should add a lot of extra value.[11,12] One example of an interesting, visually appealing and interactive interface of representing matrices is NodeTrix[13] which, concentrates on one-mode networks (graphs) and does not include support for most of the functionality described in the following sections. From the perspective of rigorous comparison of different vertex ordering and matrix permutation algorithms, the reader is referred to a recent study by Mueller et al.[14] VME, described in this article, can be thought of as a perfect environment for such comparisons, also enabling different kinds of interaction with such matrices, in order to highlight the differences and to support the final interpretation of the results.

COMPARING THEORIES

Is the psychoanalytic theory of Freud better than the self-concept theory of Rogers? While such questions, in any discipline, may be inadequate at general level, they are of primary concern for the explorer and experimenter. Even if the experimenter does not have any



Township	Public service
1	Has one-room school
1	Does not have doctor's office
2	Has agricultural cooperative
2	Has one-room school
2	Has veterinarian's office
2	Does not have doctor's office
2	Recent land reallocation
3	Has agricultural cooperative
...	...

Data acquisition (predicates)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Highschool								1			1					
Agriculture coop		1	1	1			1					1			1	
Railroad station								1			1					
One-room school	1	1			1	1			1	1			1	1		1
Veterinary		1	1	1			1					1			1	
No doctor	1	1			1	1			1	1			1	1		1
No water supply										1				1		
Police station			1					1			1					
Land reallocation		1	1	1			1					1			1	

Discretization (binary matrix representation)

FIGURE 1 | Workflow from acquiring predicate data to visual knowledge mining.

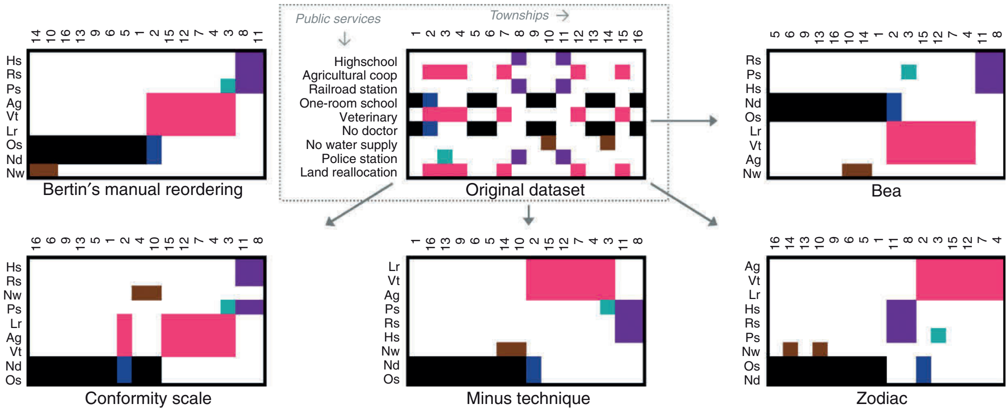


FIGURE 2 | A dataset rendered in different permutations.

preferences, there should be and efficient and objective way to analyze the differences. As stated by Kelly,[15] 'a theory may be considered as a way of binding together multitude of facts, so that one may comprehend them all at once.' Therefore, it is the facts, which allow explicit binding between different views and theories.

In the following example, the previous dataset is depicted in the top center position of Figure 2. This image also includes five other orderings from different algorithms originally meant to serve different disciplines. Technically, each of those can perform a combinatorial optimization using a different objective function, however fundamentally, all of those objective functions were tailored to emphasize as much similarities and visual clusters in the dataset as possible. These objective functions, therefore, were based on slightly different assumptions, biases and subjective beliefs.

While there may not be a single specific layout of the data that brings out the 'natural structure' of the dataset, we can identify different interesting aspects from each different view. An interested reader is referred to an extended historical review[16] in the search for the 'natural structure'. The tool presented in this article would allow comparing those different methods and approaches in order to find the natural system of things under observation.

In the view 'Bertin's manual reordering' (Figure 2) of data, VME is utilized for highlighting different subsets of data using different colors. Thereafter, the displacement of Bertin's blocks in other views of data becomes evident, allowing for a comparison of different results.

In Figure 2, magenta-colored region has the characteristics of a medium-sized rural settlement. Settlement 3 serves as a transition from a rural settlement to an urban township. The key attribute of this shift (police station) is separated by the BEA algorithm and colored in teal.

One dataset can have many different representations—in this case, permutations of rows and columns. In cellular manufacturing, 'similar' machines are placed adjacently, thereby improving the process flow,[17] by reducing worker round-trips in the factory, and minimizing the time required for transportation of materials between machines. In such a setting, rows of the matrix can represent machines while columns stand for processes, for which the machines are utilized. McCormick et al.[18] created a seriation method Bond-Energy Algorithm (BEA, Figure 2), which effectively maximizes the contiguous chunks, as clumps of data get placed next to each other.

Contiguous chunks can be interpreted as machines that should physically be placed together. If solitary bridges between chunks exist, they are, in the manufacturing workflow, usually interpreted as bottlenecks (in terms of time or material transportation).

When, from the field of social sciences, we consider social network analysis, an interesting parallel with the aforementioned manufacturing concepts can be drawn. The graph structure of a social network can be represented using an adjacency matrix, which in itself is easy to visualize. One difficulty in network analysis, though, is the calculation of a layout for a graph—as the nodes of the network do not contain any inherent ranking properties. The process of applying cellular manufacturing matrix seriation (searching for chunks) on the adjacency matrix yields a rank for each object. The chunks in the data and the bridges can, respectively, be considered as groups of friends and mediators in between, and the ranking information can be used while constructing the layout. In network analysis, a clique is an often-sought structure, which, commonly defined as an inclusive group of friends, colleagues or individuals in general, helps to identify like-minded communities sharing the same knowledge, interests or ideologies. It should be noted, though, that often the graph-theoretic definition of a clique—a group where each member is connected with every other member—cannot be exploited, because real-world data rarely exhibits perfect structures. However, an exploration of reasons behind these irregularities could be of interest, for instance, if Alice socializes with Bob and Carol, what hinders Bob from befriending Carol? VME attempts to present a supplementary instrument for analyzing networks by utilizing aforementioned visual clustering in discovery of communities and anomalies therein.

THE FUNCTIONALITY OF VME

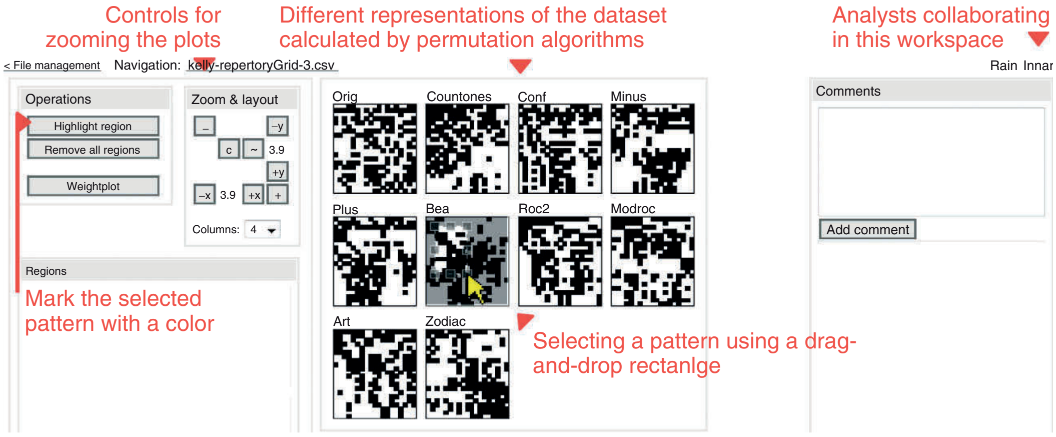
As stated in the brief discussion of seriation, research has traditionally concentrated on the best static representation of data, whereas VME enables interaction and, also, lets the user focus on a subset of interest. The functionality of the tool has been designed with the help and methodological guidance from the frameworks and taxonomies by Wehrend and Lewis,[19] Shneiderman,[20] Amar and Stasko,[21] and Amar et al.[22] that concentrate on the analytical tasks people undertake while investigating a dataset.

VME is a web-based tool that enables visualization and analysis of binary matrices. The input files are stored in a comma-separated values (CSV) format and may contain descriptive labels for rows and columns, as illustrated in the top-center plot of Figure 1. The system can be run in a web-browser supporting CSS, JavaScript, and Canvas elements.

Operation of VME is organized into workspaces. A workspace is a shared collaborative environment, which also acts as a tracker of actions taken by the user. The path of exploration and hypothesis discovery of an analyst is recorded in his 'trail of thought', which is shared across all browsers displaying the current workspace. There can be multiple workspaces of one file which is useful for following alternative trails of thought.

The workspace contains a side-by-side grid of graphical plots visualizing the dataset under different 'Permutations' (Figure 3), each of which is the result of applying one algorithm. The algorithms include:

- *orig*—stands for the original dataset;
- *countones*—our fast $O(n \log n)$ heuristic for larger matrices, based on sorting by the frequency of 'ones';
- *conf*—conformity scale; *minus*—minus technique, and *plus*—plus technique—algorithms from the Monotone Systems metaheuristic by Mullat[23] and Vyhandu.[24]
- *bea*—McCormick's BEA.[18]
- *roc2*—an enhanced rank order clustering by King et al.[25]
- *modroc*—an extension of the rank order clustering for group technology by Chandrasekharan and Rajagopalan.[26]
- *art*—a Carpenter–Grossberg neural network based clustering by Kaparthi–Suresh[27] and Kusiak–Chung.[28]


Controls for zooming the plots

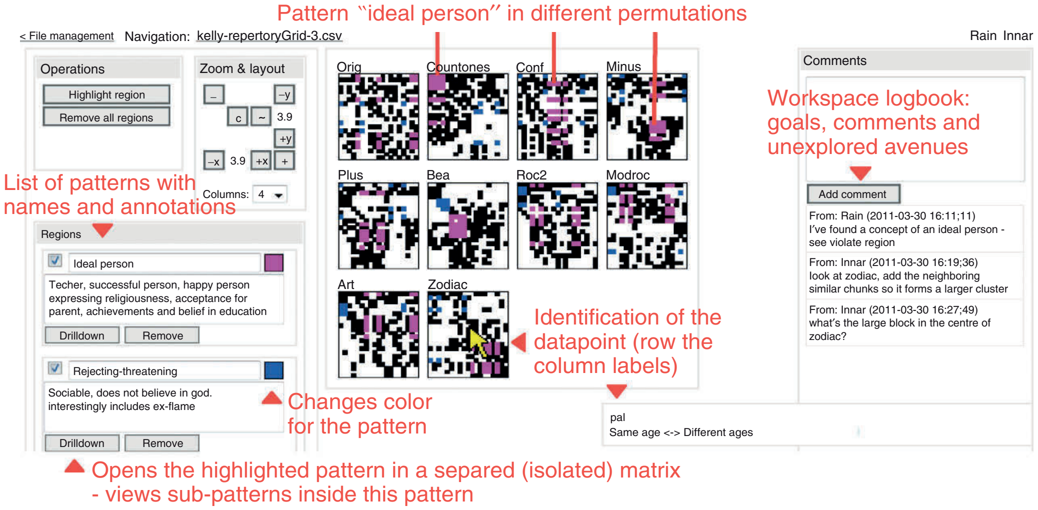
Different representations of the dataset calculated by permutation algorithms

Analysts collaborating in this workspace

Mark the selected pattern with a color

Selecting a pattern using a drag-and-drop rectanlge

FIGURE 3 | An overview of the VME workspace.



Pattern "ideal person" in different permutations

List of patterns with names and annotations

☑ Ideal person

Techer, successful person, happy person expressing religiousness, acceptance for parent, achievements and belief in education

☑ Rejecting-threatening

Sociable, does not believe in god. interestingly includes ex-flame

Changes color for the pattern

Opens the highlighted pattern in a separed (isolated) matrix - views sub-patterns inside this pattern

Identification of the datapoint (row the column labels)

pal
Same age <-> Different ages

Workspace logbook: goals, comments and unexplored avenues

From: Rain (2011-03-30 16:11;11)
I've found a concept of an ideal person - see violate region

From: Innar (2011-03-30 16:19;36)
look at zodiac, add the neighboring similar chunks so it forms a larger cluster

From: Innar (2011-03-30 16:27;49)
what's the large block in the centre of zodiac?

FIGURE 4 | Workspace with two highlighted patterns.

- *zodiac*—ideal seed method for part-family and machine-cell formation in group technology by Chandrasekharan–Rajagopalan.[29]

The system is open for extensions and new permutation algorithms can be included easily.

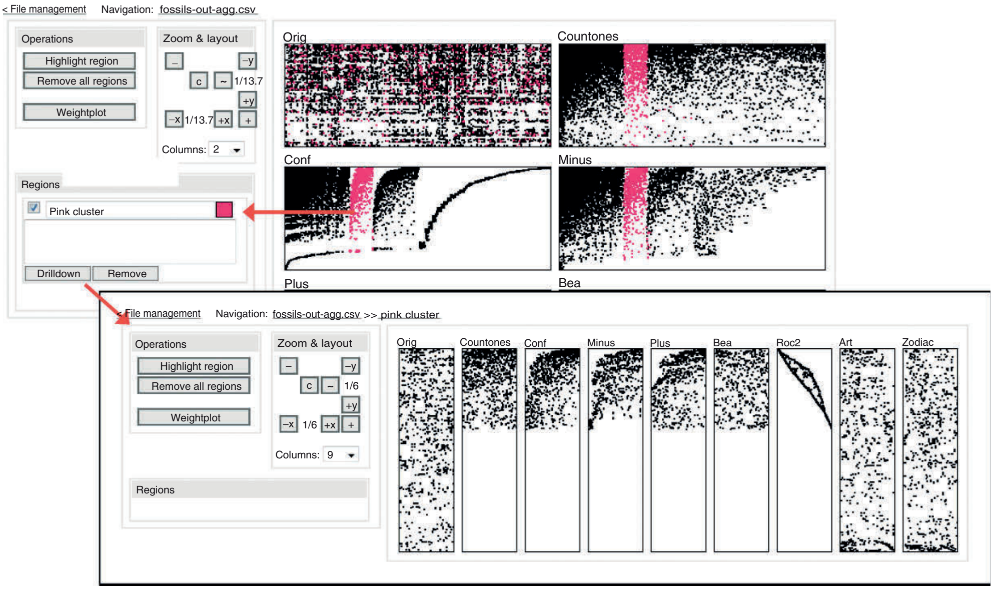
Depending on the size of the dataset and the size of the browser's viewport, a scaling factor may be necessary to fit images inside the window. In the case of large matrices, a single pixel on the screen may consist of multiple data points, similar to the information mural technique.[30]

A rectangular region of a permutation can be selected by clicking-and-dragging, as shown in the Figure 3. The selected region ('clump of points')—a subset of objects and attributes—is consequently brushed[3] with color on every plot, thereby enabling a comparison of results of different theories. In order to further identify and interpret the results, individual data points can be inspected with mouse which will result in the associated object and attribute name being displayed.

After consulting with data point names, a user can assign a descriptive name for the region. The system is designed to support easy capturing of hypotheses, for the purposes of exploration, interpretation, and reporting. Each pattern can be complemented with a detailed annotation which enables sharing the intent of the analyst instantly among fellow collaborating users. Newly identified patterns, their descriptions and explanations are delivered to all members of the workspace in a real-time fashion. A general commenting section provides a tool for driving the analysis process. It can be used as an accessible place for storing goals, unexplored alternative paths and general remarks. Figure 4 illustrates a workspace with two brushed regions along with annotations.

VME has a 'drilldown' feature that is in its nature similar to financial reporting and OLAP (see Ref 31 for an overview), in case of which transactions of a specified set of accounts are queried in detail. Drilldown can be used in order to reveal weaker patterns in the dataset (Figure 5, in pink). A weaker pattern, in our context, means a structural phenomenon of the dataset, which does not emerge clearly in the overall view; but if the structural constraints from other, more evident patterns are removed, will present interesting traits. In other words, we may say, that before drilling down we have a limited number of degrees of freedom for expressing patterns, as the rows or columns of the table cannot be split into two or more independent parts. Procedurally, the drilldown launches a new workspace that contains the results of all seriation algorithms applied only on the selected data. While considering Figure 5, we may also state, that if a set of variables and objects cause a certain pattern to emerge on many permutations, it would be of interest for the researcher to be focused on.

The time of calculation depends on the size of the dataset: for smaller datasets (less than 100 rows or columns) the results are displayed instantaneously. For moderate matrices (the number of rows and columns being between 100 and 1000), the permutations are displayed as soon as the corresponding calculation


< File management Navigation: fossils-out-agg.csv

Columns: 2

☑ Pink cluster

< File management Navigation: fossils-out-agg.csv >> pink cluster

Columns: 9

FIGURE 5 | Viewing the contents of the pink region—the results of a drilldown.

has completed. The inclusion of fast heuristic seriation methods provides user with a quick approximate overview, while more accurate albeit slower permutations are being processed. In the case of medium-sized datasets, ranging up to 5000 rows and columns, the computational complexity of seriation algorithms impacts the responsiveness of the drilldown operation, for most algorithms have time-complexity of O(n^3) or more.

We have successfully used count of ones and conformity scale methods to seriate large binary matrices of approximately 1M rows by 1k columns, although experiments show that algorithms with time-complexity beyond O(n^2) do not exhibit practically applicable execution times. As an example, the computation times for the Mammals dataset[32] described below (1597 rows and 3873 columns), vary from several seconds for simpler algorithms (*countones*, *roc*, *art*) to 20–25 min for cubic-time permutations (*plus*, *minus*, *bea*) on a virtual Intel® Pentium® G6950 2 GHz processor with 3.5 GB of memory.

The computation of permutations is handled server-side and the system is designed to launch several algorithms in parallel. The processing components can be distributed across several computers, which can include cloud-based computation services such as Amazon Elastic Compute Cloud. This architecture has relatively modest requirements for the client computer displaying the web-based user interface. However, the proposed tool is currently not suitable for visualizing large datasets because of prolonged data transfer times and limited client-side in-browser processing capabilities, being hindered mainly by browser's JavaScript engine speed.

All plots can be transformed to a spreadsheet format for detailed offline analysis. The exported files are also in the CSV format.

ILLUSTRATIVE EXAMPLES FROM DIFFERENT DISCIPLINES

In order to put forth two exemplary applications of working with VME, we are firstly going to turn to the field of psychology. The dataset depicted in Figure 6 is a classical example from Kelly's theory of Personal Constructs.[15] We are considering an interviewing technique called repertory grid, which allows a psychotherapist to identify semantic constructs of an interviewee by coming to a consensus on a common set of concepts. In order to construct the matrix, the psychotherapist enumerates a set of individuals which will range from those with concrete roles, like a mother or a friend, to abstract individuals, possessing certain values, like, for example, an ethical person. The interviewee is asked to formulate his personal constructs that are to be assigned to the figures. A construct is a contrasting or sometimes discordant pair of terms (as perceived by the

Figures																			Constructs	
Threatening person	Pitied person	Self	Mother	Happy person	Rejecting person	Ex-pal	Pal	Sister	Rejected teacher	Ethical person	Spouse	Brother	Attractive person	Boss	Ex-flame	Father	Successful person	Accepted teacher	Emergent pole	Implicit pole
			√	√						√		√		√			√	√	Achieved a lot	Hasn't achieved a lot
√	√	√	√	√	√		√		√	√		√	√	√	√	√	√	√	Believe in higher education	Not believing in too much education
					√			√							√				Don't like other people	Like other people
√		√		√	√		√		√	√	√	√	√					√	Think alike	Think differently
√		√			√	√		√			√	√	√			√	√	√	Not athletic	Athletic
			√								√		√		√				Both girls	A boy
		√	√	√			√	√	√	√		√	√			√	√	√	Both have high morals	Low morals
√	√				√	√	√	√							√				Both friends	Not friends
				√			√		√			√						√	Understand me better	Don't understand at all
				√				√	√			√		√			√		Same age	Different ages
				√			√		√	√				√			√	√	More understanding	Less understanding
											√		√		√				Both girls	Not girls
√				√	√										√		√		Don't believe in god	Very religious
							√	√	√			√		√			√	√	Higher education	No education
√	√		√	√	√							√		√				√	More sociable	Not sociable
√		√	√	√			√	√	√	√	√					√	√	√	More religious	Not religious
							√	√	√			√					√	√	Same sort of education	Completely different education
		√	√	√			√	√	√	√	√			√		√	√	√	Parents	Ideas different
	√		√	√				√	√	√				√		√	√	√	Believe the same about me	Believe differently about me
		√					√		√	√		√	√		√		√		Both appreciate music	Don't understand music
			√								√		√		√				Both girls	Not girls
			√						√	√		√		√		√		√	Teach the right thing	Teach the wrong thing

FIGURE 6 | An example repertory grid.

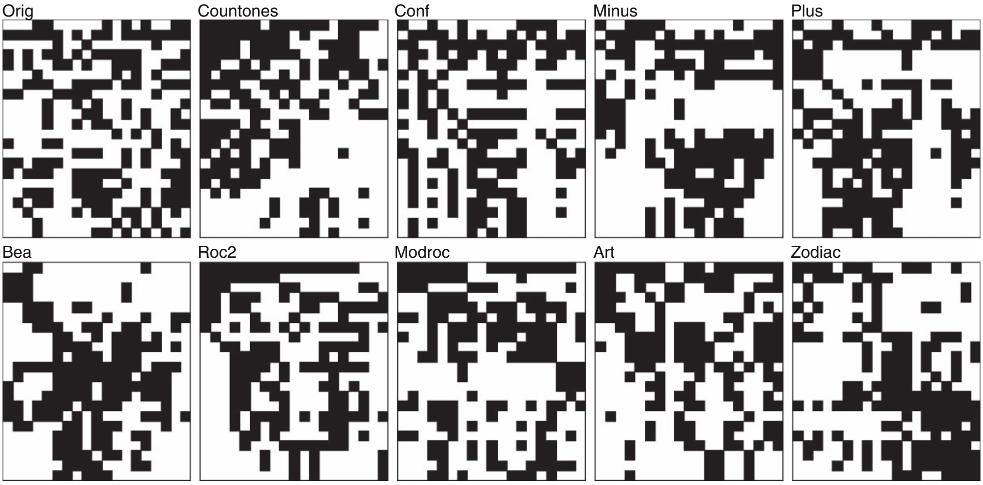


FIGURE 7 | Repertory grid visualized in multiple permutations.

interviewee), such as 'understanding person—ignorant person'. However, as the interviewee may have several meanings for the term 'understanding', it is up to the psychotherapist to identify the sets of opposite pairs—for example, to distinguish between 'understanding—ignorance' and 'understanding—disagreement'. The constructs will thereafter be applied to individuals, starting from concrete persons—family members and loved ones—and moving on to more abstract images of a threatening, attractive, or happy person. For a data analyst, Kelly's repertory grid technique can be considered as an interesting attribute discovery process for objects under investigation.

An example of the resultant table is depicted in Figure 6 which is fed into VME for seriation and visualization. In Figure 7, each tick of the grid is rendered with a black dot.

As we did in the case of Bertin's townships (Figure 2), we are now going to work with multiple

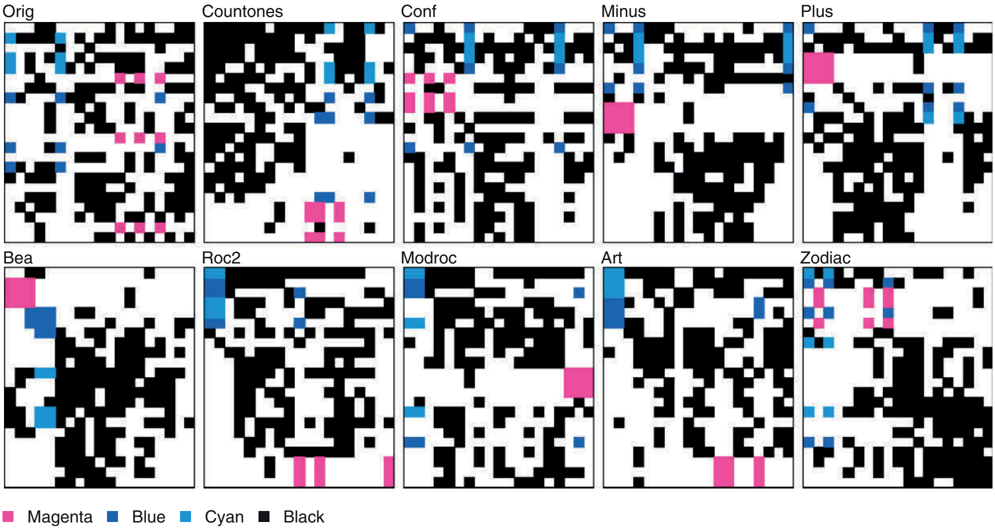


FIGURE 8 | Two highlighted patterns: magenta for a concept of loveable woman and blue for a rejecting-threatening person.

permutations. We will try to highlight relevant information therein, as relatedness is manifested in the adjacent placements by seriation algorithms.

In the *bea* permutation, the top-left corner contains two connected clumps (colored in magenta and blue in Figure 8), which in terms of cellular manufacturing can be described as sequential work cells. Consulting the attributes of the topmost cluster (in pink), it may be hypothesized to represent persons of opposite sex, who are of interest—a spouse, ex-flame, or simply an attractive person. All persons considered are female and will thereafter be marked 'loveable woman'.

Moving down from the pink cluster, the next clump is interesting. Ex-flame is placed side-by-size with an image of rejecting and threatening person. Constructs that are manifested by these roles are sociability, lack of belief in God and friendship with the interviewee. This cluster is marked with blue color in Figure 8. However, other permutations suggest including additional constructs for better comprehension of the concept 'rejecting person'.

In the *roc2* permutation, the aforementioned properties lie in a contiguous clump in the top-left corner. *Roc2* aims to solve the same cellular manufacturing problem as *bea*, however, the algorithm tries to organize the matrix in a block-diagonal form.[25] By expanding the blue-colored region, three additional descriptions of a rejecting person can be found: belief in a higher education, thinking like the interviewee and not being athletic. This operation of expanding a cluster in a parallel permutation, however well justified by the optimization algorithm's similarity measure, seems hard to vindicate in the psychological context. However, given the emotional and irrational nature of personal constructs, these controversial clusters provide a topic for a further interview.

On the *bea* permutation (Figure 9) a third, fairly solid block—that is located in the middle of the matrix—has been colored in violet. The individuals that take part in this pattern are successful person, happy person, and teacher. Looking at the represented constructs, achievements, belief in higher education, high morality, religiousness, and acceptance as a parenting role, are involved. Thus this person could be conceptualized as an 'ideal person'.

We have so far considered two processes—one being clarifying (by building roles and constructs) and the other being interpreting (e.g., by considering the permutation matrices). These can, however, be reciprocal. As the solid block mentioned above contains one inconsistency, the psychotherapist can now further investigate, whether great achievements are a prerequisite for someone to be considered a rejected teacher—as a correction would suggest.

By further trying to expand the current contiguous regions, we turn to the *zodiac* plot. The nearest similar blocks (in plum) can be attached to the violet region. This addition expands the interviewee's concept of an ideal person with two roles: mother and an ethical person. Note that this large cluster is unobtainable with a single rendering of the matrix, but in turn is seemingly justifiable in the context of our interpretation.

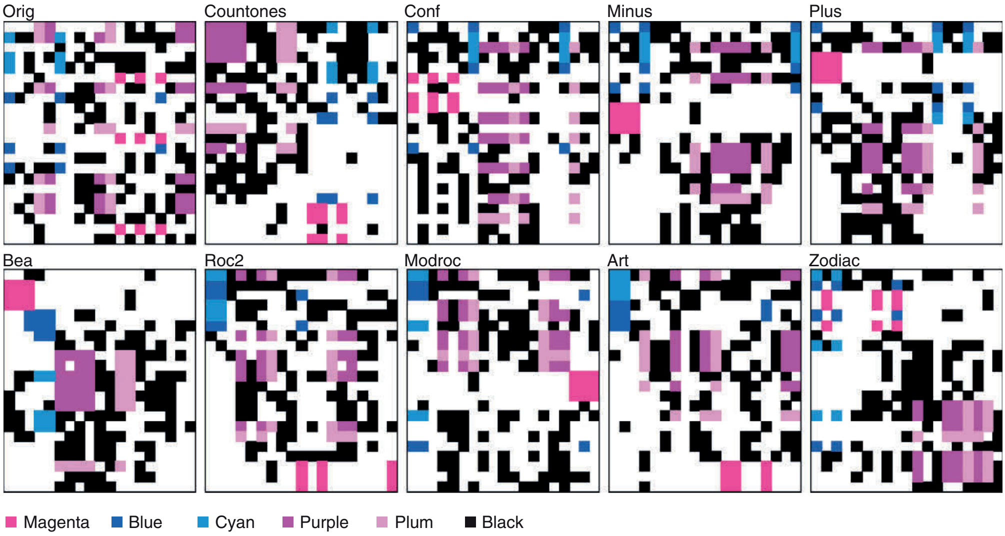
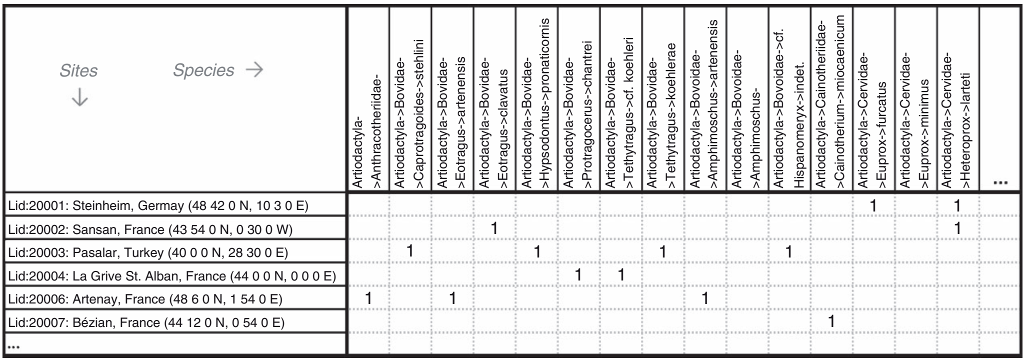


FIGURE 9 | Violet and plum regions representing a concept of an ideal person.

Secondly, we are going to consider an example from paleontology. The Neogene of the Old World[32] database contains taxa of land mammals in various localities across Europe. The original dataset includes a number of attributes: the taxon of the finding down to the species level, the geographical coordinates of the locality, an estimate on the size and the body mass of an animal, an assessment of the dietary habits of an animal, etc. (Figure 10).

The aggregated fossil matrix is much larger than the previous examples we have considered: roughly 3800 species mark the columns, while 1500 sites are on the rows of the matrix. Judging by the dimensions of the table, an investigative strategy of studying the findings of a certain species or a comparison of taxa between two sites would be too time-consuming. However, as visual clusters tend to be formed as a result of seriation, the aforementioned brush and drill-down techniques can aid in dividing the data table into smaller, clear-cut parts. A high variety of seriation algorithms implemented grants additional freedom: albeit certainly not being rigorously provable, the experience shows that at least one of the permutations usually yields some discriminating clusters.

In Figure 11, the most clear-cut seriation results are visible on *conf* and *plus* permutations (less tangibly on *minus*). The most typical findings are located in the corner of the table (in black), dissimilar species, in terms of distribution amongst localities, are colored.



Sites ↓ Species →	Artiodactyla->Anthracotheriidae-	Artiodactyla->Bovidae->Caprotragoides->stehlini	Artiodactyla->Bovidae->Eotragus->artenensis	Artiodactyla->Bovidae->Eotragus->clavatus	Artiodactyla->Bovidae->Hypsodontus->pronaticornis	Artiodactyla->Bovidae->Protragocerus->chantrei	Artiodactyla->Bovidae->Tethytragus->cf. koehleri	Artiodactyla->Bovidae->Tethytragus->koehlerae	Artiodactyla->Bovidae->Amphimoschus->artenensis	Artiodactyla->Bovidae->Amphimoschus-	Artiodactyla->Bovidae->cf. Hispanomeryx->indet.	Artiodactyla->Cainotheriidae->Cainotherium->miocaenicum	Artiodactyla->Cervidae->Euprox->furcatus	Artiodactyla->Cervidae->Euprox->minimus	Artiodactyla->Cervidae->Heteroprox->larteti	...
Lid:20001: Steinheim, Germay (48 42 0 N, 10 3 0 E)													1		1	
Lid:20002: Sansan, France (43 54 0 N, 0 30 0 W)				1											1	
Lid:20003: Pasalar, Turkey (40 0 0 N, 28 30 0 E)		1			1			1			1					
Lid:20004: La Grive St. Alban, France (44 0 0 N, 0 0 0 E)						1	1									
Lid:20006: Artenay, France (48 6 0 N, 1 54 0 E)	1		1						1							
Lid:20007: Bézian, France (44 12 0 N, 0 54 0 E)												1				
...																

FIGURE 10 | Findings of mammal species by location, a subset.

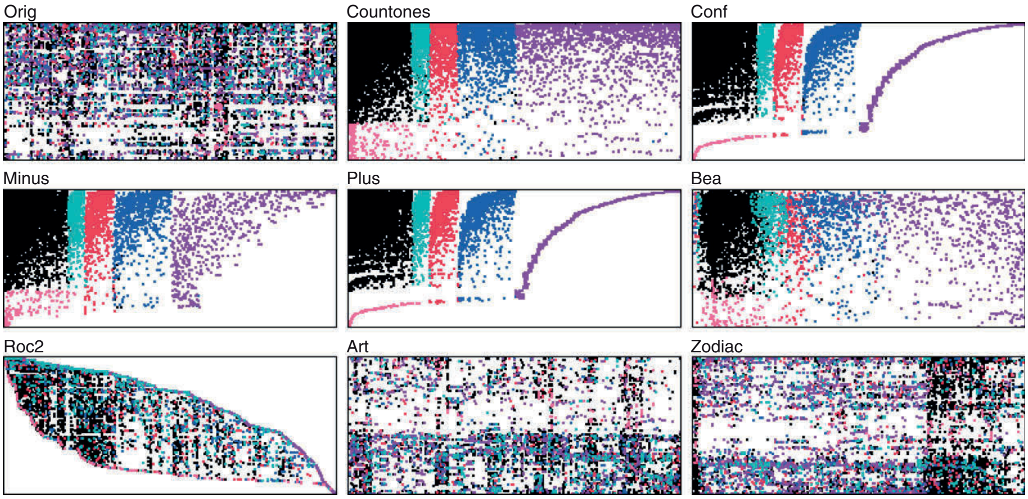


FIGURE 11 | A visualization of the fossils dataset, rows represent sites, columns mark species.

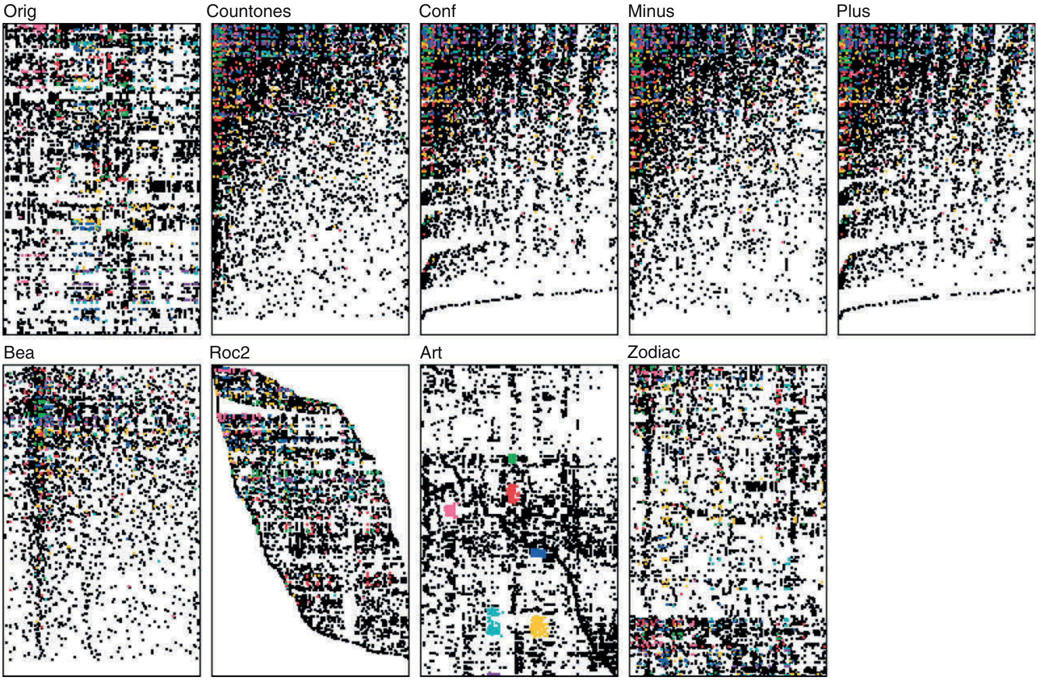


FIGURE 12 | Drilldown of the black dense region. Art and zodiac permutations introduce clumps that were not detectable on the original matrix.

According to the Monotone Systems metaheuristic algorithms (*conf*, *plus*, *minus*), the dense black part contains the strongest influencers. Drill-down operation on that part reveals a fern-like clustering on the submatrix as well (Figure 12, plots *conf* and *plus*), thus constituting a fractal-like structure. As the submatrix has more degrees of freedom, cellular manufacturing algorithms art and zodiac are able to produce fairly solid rectangular clumps.

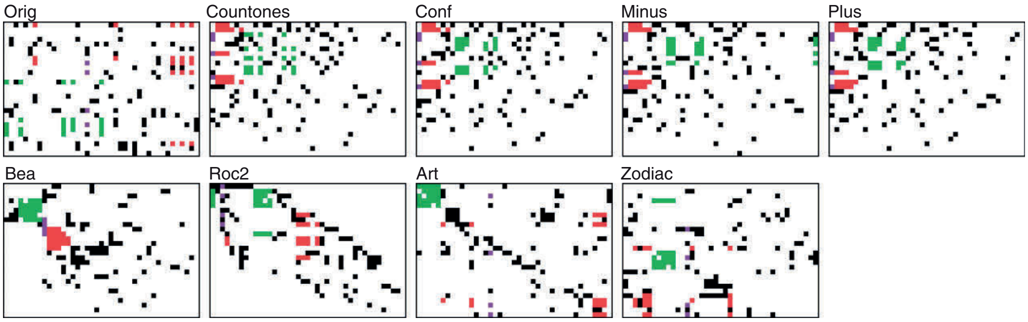


FIGURE 13 | Two sites of insectivores and rodents in Germany and France.

Figure 13 represents the blue-colored region of Figure 11. On the *bea* permutation we see clusterings with familiar shape (here colored in green and red), which are findings of several rodents and insectivores in France and Germany, the similarity of which should further be investigated. The purple line connecting the two clumps (in cellular manufacturing terms, a bottleneck), is an emphasized species (family *Soricidae*, with a common name shrew), being present at both sites.

Returning to the initial dataset in Figure 11, the loosely organized subset—in violet—can be considered as follows. The diagonal in the *plus* permutation features small connected blocks instead of a smooth curve—a hint that the datapoints can be divided into disjoint clusters of species across all sites. Drill-down of the violet region—in Figure 14—confirms the suggestion as several permutations present 5–6 disjoint clusters—a subject of discussion with fellow analysts.

CONCLUSION

This article presented a collaborative exploratory data analysis tool, VME, to analyze and compare different views of the same data. VME can also be considered as a tool to interact with seriation literature reviews and comparisons. There are a lot of tools for software visualization and information visualization, but, so far, no tools for comparing different scientific results, theories, assumptions, and biases. As an example, a bottleneck in cellular manufacturing seriation result may stand for a mediator in social network analysis.

VME also records the analytical activities and allows for web-based collaboration, as multiple users can operate on the same workspace, share comments and details of the exploration process.

We also emphasize collaboration on two different levels. First, that concerning the technical aspect of several people working simultaneously in a web-based environment. Second, allowing people

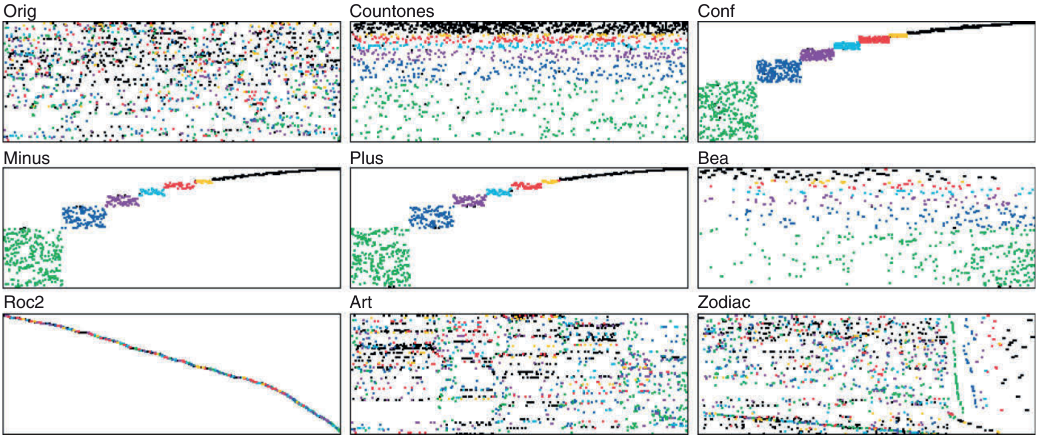


FIGURE 14 | Details of the comet-tail violet pattern in the initial matrix.

from different disciplines to work with the same dataset and to share a set of patterns, whereas maintaining their disciplines' traditional view.

In this article, VME has been applied to three datasets—those of Bertin's classical example of townships,[9] the theory of personal constructs in psychology[15] and European land mammals in paleontology.[32] Several interesting avenues of future research can be identified. Regardless of the analytics topic and task, one might be interested in logging the process of investigative data analysis. From such logs, when reasonably annotated and structured, it might be possible to extract a metaprocess for data exploration and collaborative data exploration. Certainly, a creative investigator can not be modeled, but most of the methodological repetitive steps a regular data investigator takes, could help to build a data mining, information visualization and interaction methodology using a bottom-up strategy.

VME is available as an open source project on SourceForge[33] for research purposes.

REFERENCES

1. Liiv I. Seriation and matrix reordering methods: an historical overview. *Stat Anal Data Mining* 2010, 3:70–91.
2. Inhelder B, Piaget J. *The Early Growth of Logic in the Child*. London: Routledge & Kegan Paul; 1964.
3. Becker RA, Cleveland WS. Brushing scatterplots. *Technometrics* 1987, 29: 127–142.
4. Belknap RE. *The List: The Uses and Pleasures of Cataloguing*. New Haven: Yale University Press; 2004.
5. Knight D. A Real alternative to spreadsheets. *Proc. of 2-nd Int. Symposium on Spreadsheet Risks, EUSPRIG2001*. Amsterdam, Holland, 2001.
6. Abraham R. End-User software engineering in the spreadsheet paradigm, Doctoral Thesis, Oregon State University, 2007.
7. Lenstra JK. Clustering a data array and the traveling salesman problem. *Oper Res* 1974, 22:413–414.
8. Niermann S. Optimizing the ordering of tables with evolutionary computation. *Am Stat* 2005, 59:41–46.
9. Bertin J. *Graphics and Graphic Information Processing* (Translated by Berg WJ and Scott P). Berlin: Walter de Gruyter; 1981.
10. Buchta C, Hornik K, Hashler M. Getting things in order: an introduction to the R package seriation. *J Stat Softw* 2008, 25:1–34.
11. Yi JS, Kang Y, Stasko JT, Jacko JA. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Trans Visual Comp Graph* 2007, 13:1224–1231 (Paper presented at InfoVis'07).
12. Pike WA, Stasko JT, Chang R, O'Connell TA. The science of interaction. *Inform Visual* 2009, 8:263–274.
13. Henry N, Fekete J, McGuffin MJ. NodeTrix: a hybrid visualization of social networks. *IEEE Trans Visual Comp Graph* 2007, 13:1302–1309.
14. Mueller C, Martin B, Lumsdaine A. A comparison of vertex ordering algorithms for large graph visualization. *Proceedings of Asia-Pacific Symposium on Visualization*. Sydney, Australia, 2007, 141–148.
15. Kelly GA. *The Psychology of Personal Constructs*. New York: Norton; 1955.
16. Sokal RR, Sneath PHA. *Principles of Numerical Taxonomy*. San Francisco: W. H. Freeman; 1963, 11.
17. Burbidge JL. Production flow analysis. *Prod Eng* 1963, 42: 742–752.
18. McCormick WT, Schweitzer PJ, White TW. Problem decomposition and data reorganization by a clustering technique. *Oper Res* 1972, 20:993–1009.
19. Wehrend S, Lewis C. A problem-oriented classification of visualization techniques. *Proceedings of VIS'90* 1990, 139–143.
20. Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. *Proceedings of 1996 IEEE Conference on Visual Languages*, 1996, 336–343.
21. Amar R, Stasko J. A knowledge-task based framework for design and evaluation of information visualizations. *Proceedings of InfoVis 2004*, 2004, 143–149.
22. Amar R, Eagan J, Stasko J. Low-level components of analytic activity in information visualization. *Proceedings of IEEE InfoVis'05*, Minneapolis, 2005, 111–117.
23. Mullat JE. Extremal subsystems of monotonic systems I. *Autom Remote Cont* 1976, 37:758–766.
24. Vyhandu L. Some methods to order objects and variables in data systems. *Trans of Tallinn University of Technology* 1980, 482:43–50.
25. King JR, Nakornchai V. Machine-component group formation in group technology: review and extension. *Int J Prod Res* 1982, 20:117–133.
26. Chandrasekharan MP, Rajagopalan R. MODROC: an extension of rank order clustering for group technology. *Int J Prod Res* 1986, 24:1221–1233.
27. Kaparthi S, Suresh NC. Machine-component cell formation in group technology: a neural network approach. *Int J Prod Res* 1992, 30:1353–1367.
28. Kusiak A, Chung YK. GT/ART: using neural network to form machine cells. *Manufact Rev* 1991, 4:293–301.
29. Chandrasekharan MP, Rajagopalan R. ZODIAC—an algorithm for concurrent formation of part-families and machine-cells. *Int J Prod Res* 1987, 25:835–850.

30. Jerding DF, Stasko JT. The information mural: a technique for displaying and navigating large information spaces. *IEEE Trans Visual Comp Graph* 1998, 4:257–271

31. Codd EF, Codd SB, Salley CT. *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate*. San Jose: Codd & Date; 1993.

32. Fortelius M. Neogene of the Old World Database of Fossil Mammals (NOW public release 030717). University of Helsinki, 2003. Available at: http://www.helsinki.fi/science/now. (Accessed July 2, 2011).

33. Visual Matrix Explorer source code. Available at: http://sourceforge.net/projects/vismatexplorer. (Accessed July 2, 2011).

APPENDIX D

Übi, J.; Liiv, I. (2010). A Review of Student Churn in the Light of Theories on Business Relationships. The Third International Conference on Educational Data Mining, EDM2010. Pittsburg, USA:, 2010, 329 – 330

A Review of Student Churn in the Light of Theories on Business Relationships

Jaan Ubi and Innar Liiv
Jaanbi@staff.ttu.ee
Innar.Liiv@ttu.ee
Department of Informatics, Tallinn University of Technology (TTU)

Abstract. The goal of this review is to use business theories in student retention research, which has so far been informed by economics, organizational behavior, psychology, sociology. Relationships in business networks are compared to these between students and universities, putting forth relevant characteristics for student churn/retention research. Theories regarding a taxonomy of customer churn, its determinants and consequences are also viewed in this context and implications for educational data mining (DM) are put forth.

The relationship between entities is traditionally either following a strict hierarchical fiat (HF) – if the parties belong to the same organization –, or it's essentially an arm's length transaction (ALT) – if there is a market relationship between the entities. Recent business theories describe a more nuanced reality. In the light of wider changes in research, today's corporations are found to be heterarchical; other authors speak of autonomous strategic initiatives that take place in corporations violating the principles of hierarchy. The relationships between companies are neither of ALT type. The works of Uppsala school [1] show that not only is a corporation essentially a network of units (as is elsewhere described both from multinational corporation (MNC) and its subsidiary perspective), but it is also embedded in a business network of its own. The picture is further complicated by the individualism-collectivism dimension. Big MNCs are comprised of internal markets (within one firm!) with an ongoing internal competition for world product mandates, centers-of-excellence, etc. between the subsidiary units; whereas the supply chain relationships that a firm belongs to, have been described as coevolving systems. Concentrating our attention on the relationships in business networks, we see that two firms gradually increase their commitment, as they do business with each other. A process of learning about each other's capabilities, needs and strategies takes place, as well as a formation of routines for undertaking transactions. Sides adapt to each other incrementally. Knowledge transfer is inherently present – with organizational learning taking place – the results being often tacit and intangible.

The relationship between a student and the university varies on the HF-ALT dimension. A student can simply purchase single classes from the Open University; one may be a full time student with an opportunity to call the university his *alma mater* after graduation; within the university's administrative framework, the studies can also in part be paid for by giving consults to one's peers; during the post graduate studies becoming a teaching assistant and teaching simultaneously with the studies is even more common; and finally – it is a goal of universities to populate the ranks of its faculties with the best graduates, in which case the student would administratively become a part of the organization. Entering studenthood comprises of overcoming various entry barriers. The curriculum is substantially different from that of a high school and the university studies are qualitatively harder as well – as the amount of independent work is greater, the tempo in the classes faster and as in some universities general courses can be amongst the most difficult in the undergraduate curricula. At the same students have to learn scheduling, budgeting, develop their EQ and career. The steep entry barriers underline the commitment it takes to enter the university relationship. Therefore sunk costs are formed, which are reflected in the fact that student churn lessens considerably later on. The mutual commitment and adaption is evident in the following: the student will be able to pursue further career goals after graduation; the student becomes familiar with the university life, procedures and administrative system; the youth helps to keep the university abreast of times; the university assesses its employees based on student feedback. And perhaps most importantly – the students adapt to the university and their academic mentor's profile. For both sides knowledge transfer and (organizational) learning ensue and as ties with the industry are created, so are the intangible assets.

Student churn taxonomy as a basis for further work

Customer churn is the propensity of customers to cease doing business. The cost of acquiring new customers is many times higher than that of retaining the existing ones. Customer (monetary) lifetime value(CLV) has been linked to customer tenure and is also something that previous research has considered in educational settings, although misinterpreting the NPV term**.** The most important determinants of customer churn are switching costs, satisfaction and future usage. The switching costs are transaction costs (while transferring from one university to another), learning costs (differences in curricula) and artificial costs (a scholarship keeping a student at the university). It can be inferred from the discussion above that satisfaction is a key determinant of student churn and that future usage plays an important role in the relationship. Bringing a parallel from business, according to a taxonomy [2], involuntary external churn occurs in the case of exmatriculation or accrued debt; voluntary external churn in case of relocation, switch to another university or alternative career path, for family considerations (this type can furher be classified as either deliberate or incidental according to whether the locus of origin lies with the student); internal churn takes place inside the university framework – such as moving between full time and part time study plan; churn can furthermore be either customer or competitor initiated, the one at hand being mainly the former.

This review has resulted in an identification of a need to develop a classification model for several student churn definitions, which will additionally be informed by the student-university relationship information. Performance, demograhpic, high school and satisfaction data is available. DM and visualization will use seriation, matrix reordering, clustering and feature selection. As research has combined social network analysis with DM, student network data (the classes taken with the peers; the supervisors; the specialization; but also in the mid-term the facebook data) will be utilized. The retention effort is pinpointed by „quantifying" the students (CLV). In Estonia, projects analyzing the global effectiveness of students allocation, further reducing churn, are possible.

References

[1] Forsgren, M. *Managing the embedded multinational,* 2005. UK: Edward Elgar.

[2] Mattison, R. *The telco churn management handbook.*, 2005 Illinois: XiT Press.

CURRICULUM VITAE

Personal Data

Name:	Jaan Übi
Date of birth:	27.09.1977
Place of birth:	Estonia
Citizenship:	Estonian

Contact data

Phone:	+37256482063
E-mail:	jaan.ubi@ttu.ee
Blog:	jaanbi.blogspot.com

Education

2008–...	Ph.D. student, Department of Informatics, Tallinn University of Technology
2005–2008	M.Sc. in Informatics, TUT
1995–2003	B.A. in Business Administration, University of Tartu
1992–1995	High school diploma, Tallinn's Reaalkool
–1992	Primary school diploma, Tallinn's Reaalkool

Career

2012-2012	Part-time lecturer, University of Applied Sciences
2007–	Part-time lecturer, TSEBA of Tallinn University of Technology
2005– …	Assistant Lecturer, Tallinn University of Technology
2004–2005	Engineer, Tallinn University of Technology

Academic Degree

Master of Science in Informatics, TUT
"E-learning course "Optimization Modeling with Spreadsheets""

Sabatical Semester

Visiting Scholar during a sabbatical semester at the University of Georgia, USA – 01/2011 – 07/2011

Membership in organizations

Charter member of International Educational Data Mining Society
The Institute for Operations Research and Management Sciences (INFORMS)
Estonian Mathematical Society
German Operations Research Society - Gesellschaft für Operations Research

Lecturing in China and the US

Shanghai University of Science and Technology
China University of Petroleum
University of Science and Technology Beijing
Macau University of Science and Technology, twice
Salisbury University, Maryland

Developing contacts with other universities – visits of foreign scientists to Estonia, organized and carried through

Assistant Professor Juan Cui of the University of Georgia/ the University of Nebraska Lincoln 05/2012-06/2012

Department Chair Kathleen Wright and Professor of Practice Paula Morris of Salisbury University, Maryland 05/2013-06/2013

Faculty Chair Ah Chung Tsoi of Macau University of Science and Technology 07/2013

Yahoo! Fellow in Residence Kalev Leetaru of Georgetown University in Washington DC 12/2013

Number of students taught (2004-2014): 1900

Teacher reports (on a scale 1..5)

or. nr.	Question	2009/2010 - S	2009/2010 - K	2010/2011 - S	2010/2011 - K	2011/2012 - S	2011/2012 - K	2012/2013 - S	2012/2013 - K
1	Presentation of course topics	4.67	4	4.27	3	3.75	4.05	4.36	4.64
2	Comprehensibility of the course	4.33	3.57	3.92	3	3.17	3.76	3.75	4.22
3	Impartial attitude to students during the course	3.33	4.38	4.31	5	4.67	4.86	4.89	4.8
4	Use of modern teaching methods	4.67	4.71	4.58	5	4	4.81	4.79	4.79
	gallup avg.	**4.25**	**4.17**	**4.27**	**4**	**3.9**	**4.37**	**4.45**	**4.61**
	ans. students	2	7	34	2	6	11	14	43
	dec. students	91	92	168	14	125	122	89	111
	ans. students(%)	2.20%	7.61%	20.24%	14.29%	4.80%	9.02%	15.73%	38.74%

CURRICULUM VITAE (IN ESTONIAN)

Isikuandmed

Nimi	Jaan Übi
Sünniaeg:	27.09.1977
Sünnikoht:	Eesti
Kodakondsus:	Eesti

Kontaktandmed

Telefon	+37256482063
E-mail:	jaan.ubi@ttu.ee
Blog:	jaanbi.blogspot.com

Hariduskäik

2008–...	doktorant, informaatikainstituut, Tallinna Tehnikaülikool
2005–2008	tehnikateaduste magister, informaatika, Tallinna Tehnikaülikool
1995–2003	sotsiaalteaduste bakalaureus, ettevõttemajanduse eriala, Tartu Ülikool
1992–1995	Keskkooli diplom, Tallinna Reaalkool
–1992	Põhikooli diplom, Tallinna Reaalkool

Teenistukäik

2012–2012	Tunnitasuline õppejõud, Tallinna Tehnikakõrgkool
2007–	Tunnitasuline õppejõud, TSEBA, Tallinna Tehnikaülikool
2005– …	assistent, Tallinna Tehnikaülikool
2004–2005	insener, Tallinna Tehnikaülikool

Akadeemiline kraad

Tehnikateaduste magister, informaatikainstituut, TTÜ “E-õppe kursus "Optimeerimismodelleerimine tabeliprogrammidega"”

Vaba semester

Külalisteadlane, the University of Georgia, USA – 01/2011 – 07/2011

Liikmelisus organisatsioonides

International Educational Data Mining Society
The Institute for Operations Research and Management Sciences (INFORMS)
Estonian Mathematical Society
German Operations Research Society - Gesellschaft für Operations Research

Loengute pidamine Hiinas ja USAs

Shanghai University of Science and Technology
China University of Petroleum
University of Science and Technology Beijing
Macau University of Science and Technology, kaks korda
Salisbury University, Maryland

Kontaktide loomine välisülikoolidega – organiseeritud välisteadlaste visiidid Eestisse

Assistant Professor Juan Cui, the University of Georgia/ the University of Nebraska Lincoln 05/2012-06/2012

Department Chair Kathleen Wright ja Professor of Practice Paula Morris, Salisbury University, Maryland 05/2013-06/2013

Faculty Chair Ah Chung Tsoi, Macau University of Science and Technology 07/2013

Yahoo! Fellow in Residence Kalev Leetaru, Georgetown University, Washington DC 12/2013

Õpetatud tudengeid(2004-2014): 1900

Tudengite hinded(skaalal 1..5)

või. nr.	Küsimus	2009/2010 - S	2009/2010 - K	2010/2011 - S	2010/2011 - K	2011/2012 - S	2011/2012 - K	2012/2013 - S	2012/2013 - K
1	Kursuse teema esitlus	4.67	4	4.27	3	3.75	4.05	4.36	4.64
2	Kursuse arusaadavus	4.33	3.57	3.92	3	3.17	3.76	3.75	4.22
3	Erapooletu suhtumine tudengitesse	3.33	4.38	4.31	5	4.67	4.86	4.89	4.8
4	Kaasaegsete õpetamismeetodite ja vahendite kasutamine	4.67	4.71	4.58	5	4	4.81	4.79	4.79
	gallupi keskmine	**4.25**	**4.17**	**4.27**	**4**	**3.9**	**4.37**	**4.45**	**4.61**
	vastanud tudengeid	2	7	34	2	6	11	14	43
	deklareerinud tudengeid	91	92	168	14	125	122	89	111
	vastanute protsent(%)	2.20%	7.61%	20.24%	14.29%	4.80%	9.02%	15.73%	38.74%

DISSERTATIONS DEFENDED AT TALLINN UNIVERSITY OF TECHNOLOGY ON *INFORMATICS AND SYSTEM ENGINEERING*

1. **Lea Elmik**. Informational Modelling of a Communication Office. 1992.

2. **Kalle Tammemäe**. Control Intensive Digital System Synthesis. 1997.

3. **Eerik Lossmann**. Complex Signal Classification Algorithms, Based on the Third-Order Statistical Models. 1999.

4. **Kaido Kikkas**. Using the Internet in Rehabilitation of People with Mobility Impairments – Case Studies and Views from Estonia. 1999.

5. **Nazmun Nahar**. Global Electronic Commerce Process: Business-to-Business. 1999.

6. **Jevgeni Riipulk**. Microwave Radiometry for Medical Applications. 2000.

7. **Alar Kuusik**. Compact Smart Home Systems: Design and Verification of Cost Effective Hardware Solutions. 2001.

8. **Jaan Raik**. Hierarchical Test Generation for Digital Circuits Represented by Decision Diagrams. 2001.

9. **Andri Riid**. Transparent Fuzzy Systems: Model and Control. 2002.

10. **Marina Brik**. Investigation and Development of Test Generation Methods for Control Part of Digital Systems. 2002.

11. **Raul Land**. Synchronous Approximation and Processing of Sampled Data Signals. 2002.

12. **Ants Ronk**. An Extended Block-Adaptive Fourier Analyser for Analysis and Reproduction of Periodic Components of Band-Limited Discrete-Time Signals. 2002.

13. **Toivo Paavle**. System Level Modeling of the Phase Locked Loops: Behavioral Analysis and Parameterization. 2003.

14. **Irina Astrova**. On Integration of Object-Oriented Applications with Relational Databases. 2003.

15. **Kuldar Taveter**. A Multi-Perspective Methodology for Agent-Oriented Business Modelling and Simulation. 2004.

16. **Taivo Kangilaski**. Eesti Energia käiduhaldussüsteem. 2004.

17. **Artur Jutman**. Selected Issues of Modeling, Verification and Testing of Digital Systems. 2004.

18. **Ander Tenno**. Simulation and Estimation of Electro-Chemical Processes in Maintenance-Free Batteries with Fixed Electrolyte. 2004.

19. **Oleg Korolkov**. Formation of Diffusion Welded Al Contacts to Semiconductor Silicon. 2004.

20. **Risto Vaarandi**. Tools and Techniques for Event Log Analysis. 2005.

21. **Marko Koort**. Transmitter Power Control in Wireless Communication Systems. 2005.

22. **Raul Savimaa**. Modelling Emergent Behaviour of Organizations. Time-Aware, UML and Agent Based Approach. 2005.

23. **Raido Kurel**. Investigation of Electrical Characteristics of SiC Based Complementary JBS Structures. 2005.

24. **Rainer Taniloo**. Ökonoomsete negatiivse diferentsiaaltakistusega astmete ja elementide disainimine ja optimeerimine. 2005.

25. **Pauli Lallo.** Adaptive Secure Data Transmission Method for OSI Level I. 2005.

26. **Deniss Kumlander**. Some Practical Algorithms to Solve the Maximum Clique Problem. 2005.

27. **Tarmo Veskioja**. Stable Marriage Problem and College Admission. 2005.

28. **Elena Fomina**. Low Power Finite State Machine Synthesis. 2005.

29. **Eero Ivask**. Digital Test in WEB-Based Environment 2006.

30. **Виктор Войтович**. Разработка технологий выращивания из жидкой фазы эпитаксиальных структур арсенида галлия с высоковольтным p-n переходом и изготовления диодов на их основе. 2006.

31. **Tanel Alumäe**. Methods for Estonian Large Vocabulary Speech Recognition. 2006.

32. **Erki Eessaar**. Relational and Object-Relational Database Management Systems as Platforms for Managing Softwareengineering Artefacts. 2006.

33. **Rauno Gordon**. Modelling of Cardiac Dynamics and Intracardiac Bio-impedance. 2007.

34. **Madis Listak**. A Task-Oriented Design of a Biologically Inspired Underwater Robot. 2007.

35. **Elmet Orasson**. Hybrid Built-in Self-Test. Methods and Tools for Analysis and Optimization of BIST. 2007.

36. **Eduard Petlenkov**. Neural Networks Based Identification and Control of Nonlinear Systems: ANARX Model Based Approach. 2007.

37. **Toomas Kirt**. Concept Formation in Exploratory Data Analysis: Case Studies of Linguistic and Banking Data. 2007.

38. **Juhan-Peep Ernits**. Two State Space Reduction Techniques for Explicit State Model Checking. 2007.

39. **Innar Liiv**. Pattern Discovery Using Seriation and Matrix Reordering: A Unified View, Extensions and an Application to Inventory Management. 2008.

40. **Andrei Pokatilov**. Development of National Standard for Voltage Unit Based on Solid-State References. 2008.

41. **Karin Lindroos**. Mapping Social Structures by Formal Non-Linear Information Processing Methods: Case Studies of Estonian Islands Environments. 2008.

42. **Maksim Jenihhin**. Simulation-Based Hardware Verification with High-Level Decision Diagrams. 2008.

43. **Ando Saabas**. Logics for Low-Level Code and Proof-Preserving Program Transformations. 2008.

44. **Ilja Tšahhirov**. Security Protocols Analysis in the Computational Model – Dependency Flow Graphs-Based Approach. 2008.

45. **Toomas Ruuben**. Wideband Digital Beamforming in Sonar Systems. 2009.

46. **Sergei Devadze**. Fault Simulation of Digital Systems. 2009.

47. **Andrei Krivošei**. Model Based Method for Adaptive Decomposition of the Thoracic Bio-Impedance Variations into Cardiac and Respiratory Components. 2009.

48. **Vineeth Govind**. DfT-Based External Test and Diagnosis of Mesh-like Networks on Chips. 2009.

49. **Andres Kull**. Model-Based Testing of Reactive Systems. 2009.

50. **Ants Torim**. Formal Concepts in the Theory of Monotone Systems. 2009.

51. **Erika Matsak**. Discovering Logical Constructs from Estonian Children Language. 2009.

52. **Paul Annus**. Multichannel Bioimpedance Spectroscopy: Instrumentation Methods and Design Principles. 2009.

53. **Maris Tõnso**. Computer Algebra Tools for Modelling, Analysis and Synthesis for Nonlinear Control Systems. 2010.

54. **Aivo Jürgenson**. Efficient Semantics of Parallel and Serial Models of Attack Trees. 2010.

55. **Erkki Joasoon**. The Tactile Feedback Device for Multi-Touch User Interfaces. 2010.

56. **Jürgo-Sören Preden**. Enhancing Situation – Awareness Cognition and Reasoning of Ad-Hoc Network Agents. 2010.

57. **Pavel Grigorenko**. Higher-Order Attribute Semantics of Flat Languages. 2010.

58. **Anna Rannaste**. Hierarcical Test Pattern Generation and Untestability Identification Techniques for Synchronous Sequential Circuits. 2010.

59. **Sergei Strik**. Battery Charging and Full-Featured Battery Charger Integrated Circuit for Portable Applications. 2011.

60. **Rain Ottis**. A Systematic Approach to Offensive Volunteer Cyber Militia. 2011.

61. **Natalja Sleptšuk**. Investigation of the Intermediate Layer in the Metal-Silicon Carbide Contact Obtained by Diffusion Welding. 2011.

62. **Martin Jaanus**. The Interactive Learning Environment for Mobile Laboratories. 2011.

63. **Argo Kasemaa**. Analog Front End Components for Bio-Impedance Measurement: Current Source Design and Implementation. 2011.

64. **Kenneth Geers**. Strategic Cyber Security: Evaluating Nation-State Cyber Attack Mitigation Strategies. 2011.

65. **Riina Maigre**. Composition of Web Services on Large Service Models. 2011.

66. **Helena Kruus**. Optimization of Built-in Self-Test in Digital Systems. 2011.

67. **Gunnar Piho**. Archetypes Based Techniques for Development of Domains, Requirements and Sofware. 2011.

68. **Juri Gavšin**. Intrinsic Robot Safety Through Reversibility of Actions. 2011.

69. **Dmitri Mihhailov**. Hardware Implementation of Recursive Sorting Algorithms Using Tree-like Structures and HFSM Models. 2012.

70. **Anton Tšertov**. System Modeling for Processor-Centric Test Automation. 2012.

71. **Sergei Kostin**. Self-Diagnosis in Digital Systems. 2012.

72. **Mihkel Tagel**. System-Level Design of Timing-Sensitive Network-on-Chip Based Dependable Systems. 2012.

73. **Juri Belikov**. Polynomial Methods for Nonlinear Control Systems. 2012.

74. **Kristina Vassiljeva**. Restricted Connectivity Neural Networks based Identification for Control. 2012.

75. **Tarmo Robal**. Towards Adaptive Web – Analysing and Recommending Web Users` Behaviour. 2012.

76. **Anton Karputkin**. Formal Verification and Error Correction on High-Level Decision Diagrams. 2012.

77. **Vadim Kimlaychuk**. Simulations in Multi-Agent Communication System. 2012.

78. **Taavi Viilukas**. Constraints Solving Based Hierarchical Test Generation for Synchronous Sequential Circuits. 2012.

79. **Marko Kääramees**. A Symbolic Approach to Model-based Online Testing. 2012.

80. **Enar Reilent**. Whiteboard Architecture for the Multi-agent Sensor Systems. 2012.

81. **Jaan Ojarand**. Wideband Excitation Signals for Fast Impedance Spectroscopy of Biological Objects. 2012.

82. **Igor Aleksejev**. FPGA-based Embedded Virtual Instrumentation. 2013.

83. **Juri Mihhailov**. Accurate Flexible Current Measurement Method and its Realization in Power and Battery Management Integrated Circuits for Portable Applications. 2013.

84. **Tõnis Saar**. The Piezo-Electric Impedance Spectroscopy: Solutions and Applications. 2013.

85. **Ermo Täks**. An Automated Legal Content Capture and Visualisation Method. 2013.

86. **Uljana Reinsalu**. Fault Simulation and Code Coverage Analysis of RTL Designs Using High-Level Decision Diagrams. 2013.

87. **Anton Tšepurov**. Hardware Modeling for Design Verification and Debug. 2013.

88. **Ivo Müürsepp**. Robust Detectors for Cognitive Radio. 2013.

89. **Jaas Ježov**. Pressure sensitive lateral line for underwater robot. 2013.

90. **Vadim Kaparin**. Transformation of Nonlinear State Equations into Observer Form. 2013.

91. **Ingrid Pappel**. Development and Implementation of e-Governance Framework and Paperless Management in Local Governments. 2014.

92. **Reeno Reeder**. Improvement and Optimisation of Simulation Software for Teraherz Range Radiaton Sources Based on Quantum Well Heterostructures. 2014.

93. **Ants Koel**. GaAs and SiC Semiconductor Materials Based Power Structures: Static and Dynamic Behavior Analysis. 2014.